1 Floating point arithmetic

**Problem 1** Find the decimal representation of the rational number \( x = \frac{1}{7} \).

**Problem 2** Round the decimal representation of \( x = \frac{1}{7} \) to 9 significant figures and calculate the exact value of the absolute and relative rounding error.

**Problem 3** Find the binary representation of the real number \( x = \frac{1}{7} \).

**Problem 4** Round the binary representation of the real number \( x = \frac{1}{7} \) to 9 significant figures. Compute the exact value of the absolute and relative rounding error.

**Problem 5** Find the binary representation of \( x = \frac{3}{7} \).

**Problem 6** Find the binary expansion of \( x = \frac{3}{7} \) to seven significant figures. Compute the exact absolute and relative rounding errors.

**Problem 7** Find the trinary expansion of the real number \( x = \frac{1}{7} \).

**Problem 8** Round the trinary expansion of \( x = \frac{1}{7} \) to 6 significant figures. Compute the absolute and relative rounding errors.

**Problem 9** Find the binary representation of \( x = \frac{5}{11} \).

**Problem 10** Find the binary representation of \( x = \frac{7}{13} \).

**Problem 11** Let \( x = 0.6948 \) and \( y = 0.3171 \). Compute, by hand, to four significant figures the result of each of the following operations

\[
x + y, \quad x - y, \quad x \times y, \quad \frac{x}{y}
\]

**Problem 12** Let \( \{x_i\}_{i=1}^4 \) be given by the following table:

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( 9.508 \cdot 10^{-1} )</td>
</tr>
<tr>
<td>2</td>
<td>( 6.946 \cdot 10^{-1} )</td>
</tr>
<tr>
<td>3</td>
<td>( 4.113 \cdot 10^{-4} )</td>
</tr>
<tr>
<td>4</td>
<td>( 2.789 \cdot 10^{-4} )</td>
</tr>
</tbody>
</table>

1. Compute the sum

\[
r = \sum_{i=1}^{4} x_i
\]

using exact arithmetic.

2. Compute the sum

\[
s = (((x_1 + x_2) + x_3) + x_4),
\]

by rounding each intermediate result to four significant figures. Notice that we are adding the largest numbers first!
3. Compute the sum
\[ t = (((x_4 + x_3) + x_2) + x_1), \]
by rounding each intermediate result to four significant figures. Notice that we are adding the smallest number first!

4. Compare \( \hat{s} \) and \( \hat{t} \) to the exact result \( r \) by computing the absolute error and the relative error in each case.

1.1 Calculation of averages

Problem 13 Let \( \{(x_i, y_i)\}_{i=1}^4 \) be given by the following table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( x_i )</td>
<td>( y_i )</td>
<td>( s_i )</td>
<td>( \hat{\mu}_i )</td>
</tr>
<tr>
<td>1</td>
<td>6.178 \cdot 10^{-1}</td>
<td>2.803 \cdot 10^{-1}</td>
<td>5.127 \cdot 10^{-2}</td>
<td>5.128 \cdot 10^{-2}</td>
</tr>
<tr>
<td>2</td>
<td>5.127 \cdot 10^{-2}</td>
<td>6.241 \cdot 10^{-3}</td>
<td>6.243 \cdot 10^{-2}</td>
<td>6.243 \cdot 10^{-2}</td>
</tr>
<tr>
<td>3</td>
<td>6.241 \cdot 10^{-1}</td>
<td>2.654 \cdot 10^{-1}</td>
<td>2.657 \cdot 10^{-2}</td>
<td>2.657 \cdot 10^{-2}</td>
</tr>
<tr>
<td>4</td>
<td>2.654 \cdot 10^{1}</td>
<td>2.657 \cdot 10^{2}</td>
<td>2.657 \cdot 10^{2}</td>
<td>2.657 \cdot 10^{2}</td>
</tr>
</tbody>
</table>

1. Complete the following table:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where
\[ s_i = x_i + y_i, \quad \mu_i = \frac{s_i}{2}, \]
and \( \hat{z} \) is \( z \) rounded to four significant figures.

2. For \( i = 1, 2, 3, 4 \) determine the location of the computed value of \( \mu_i \), i.e. the number
\[ \hat{\mu}_i = \hat{\mu} \left( \frac{s_i}{2} \right), \]
with respect to the interval with endpoints \( x_i \) and \( y_i \). Is \( \hat{\mu}_i \) located to the left, to the right or does it actually lie inside the interval as we would like?

Problem 14 Problem 13 revealed the computed average \( \hat{\mu} \) of two numbers \( x \) and \( y \) does not necessarily lie in the interval between the numbers! The formula
\[ \mu = \frac{x + y}{2} = x + \frac{y - x}{2} \]
suggests a different strategy, specifically,

\[ \hat{\mu} = \text{fl} \left( x + \text{fl} \left( \frac{\text{fl}(y - x)}{2} \right) \right), \]

where \( \text{fl}(z) \) is the value of \( z \) rounded to, say, 4 significant figures. Demonstrate that this procedure is sufficient to overcome the problems you encountered in Problem 13.

### 1.2 Significant figures

**Problem 15** It is known that the real number \( x \) satisfies the inequality

\[ 0.67854 < x < 0.67855. \]

Find the value of \( x \) rounded correctly to \( k \) significant figures for \( k \) as large as possible and write the result using normalized scientific notation.

**Problem 16** It is known that the real number \( x \) satisfies the inequality

\[ 0.67854 < x < 0.67856. \]

Find the value of \( x \) rounded correctly to \( k \) significant figures for \( k \) as large as possible and write the result using normalized scientific notation.

### 1.3 The Futility of Mathematical Overkill

**Problem 17** Let \( x \) be any nonzero real number and let \( y \) be an approximation of \( x \), such that the relative error satisfies

\[ \frac{|x - y|}{|x|} \leq \delta \quad (1) \]

for some \( \delta > 0 \). Show that the absolute error satisfies

\[ |x - y| \leq \frac{\delta}{1 - \delta} |y| \]

provided that \( \delta < 1 \). Deduce further that

\[ y - \frac{\delta}{1 - \delta} |y| \leq x \leq y + \frac{\delta}{1 - \delta} |y|. \quad (2) \]

provided that \( \delta < 1 \).

**Problem 18** Let \( x, y, \) and \( z \) be real numbers such that

\[ |x - y| \leq \delta |x|, \quad |y - z| \leq \nu |y| \]
where \( \delta > 0 \) and \( \nu > 0 \). Show that

\[
|x - z| \leq (2\delta + \nu)|x|
\]

**Remark 1** Problem 18 is very simple, but the consequences are very deep as it can be used to explain the futility of mathematical overkill! Suppose we wish to compute \( x \), but the actual formula for \( x \) is so complicated that we settle for a good approximation \( y \). Typically, the quality of \( y \) improves with the amount of work we are willing to do, so when should we stop? We can never hope to compute the exact value \( y \), at best we can find the floating point representation of \( y \), i.e. \( z = \text{fl}(y) \) and that would be considered an extraordinary accomplishment! We have

\[
|x - z| \leq (2\delta + \nu)|x|
\]

where \( \delta \) reflects the quality of \( y \) as an approximation of \( x \), while \( \nu \) is the unit round off error of your computer. Obviously, there is no point in achieving \( \delta \ll \nu \) and any sane person would be extremely pleased with \( \delta \sim \nu \).

### 1.4 Sums and products

**Problem 19** Let \( u > 0 \) and define

\[
\gamma_n = \frac{nu}{1 - nu}
\]  

(3)

where \( n \in \mathbb{N} \) and \( nu < 1 \). Show that

\[
\gamma_n + \gamma_m + \gamma_n \gamma_m \leq \gamma_{n+m}
\]  

(4)

whenever \((n + m)u < 1\).

**Remark 2** Problem 19 is central to our understanding the limitations of floating point arithmetic. Please remember it for a very long time.

**Problem 20** Consider the addition of four floating point numbers \( x \), \( y \), and \( z \). Let

\[
s = ((x + y) + z) + v,
\]

where the brackets indicate the order in which the additions are performed. Show that there exists real numbers \( \theta_1^{(3)}, \theta_2^{(2)}, \) and \( \theta_3^{(1)} \), such that the computed value \( \hat{s} \) satisfies

\[
\hat{s} = s + \{x + y\} \theta_1^{(3)} + y \theta_2^{(2)} + z \theta_3^{(1)}, \quad |\theta_i^{(j)}| \leq \gamma_j,
\]

where

\[
\gamma_n = \frac{nu}{1 - nu}, \quad nu < 1,
\]
and \(u\) is the unit roundoff error.

**Problem 21** Let \(u\) denote the unit round off error and let \(\{\delta_j\}_{j=1}^n\) be a sequence of real numbers such that
\[
\forall j : |\delta_j| \leq u.
\]
Show that
\[
\left| \left(\prod_{j=1}^n (1 + \delta_j)\right) - 1 \right| \leq \gamma_n = \frac{nu}{1 - nu},
\]
provided that \(nu < 1\).

**Problem 22** Let \(\{x_j\}_{j=1}^n\) be a sequence of machine numbers and let \(s_j\) be given recursively by
\[
s_1 = x_1, \quad s_j = s_{j-1}x_j, \quad j = 2, 3, \ldots, n
\]
Show that the computed value \(\hat{s}_n\) satisfies
\[
\exists \theta : (\hat{s}_n = s_n(1 + \theta) \land |\theta| \leq \gamma_{n-1}),
\]
provided that \(nu < 1\) and that the intermediate calculations do not experience overflow or underflow.

## 2 Heron’s formula

Heron’s formula can be used to compute the area \(A\) of a triangle in the plane. Specifically, if the sides are called \(a\), \(b\), and \(c\), then
\[
A^2 = \frac{1}{16}p(p-a)(p-b)(p-c)
\]
where the central parameter \(p\) is given by
\[
p = \frac{a + b + c}{2}.
\]
This formula is correct, however it is possible to choose numbers \(a\), \(b\) and \(c\), such that the computer will return a negative value for \(A^2\). In the following problems you will explore this problem and find a way to avoid this problem.

**Problem 23** Compute the exact value of \(A^2\) for triangle with sides \(a = 0.6374\), \(b = 0.6377\), and \(c = 0.0004\) using Heron’s formula (5). Repeat the calculation, but round every intermediate result to four significant figures!

The following problem shows that the naive application of Heron’s formula does not necessarily breakdown in the sense that it returns a negative value for \(A^2\). However, who wants an algorithm which only works occasionally?

**Problem 24** Compute the exact value of \(A^2\) of a triangle with sides \(a = 0.6378\), \(b = 0.6377\) and \(c = 0.0004\) using Heron’s formula (5). Repeat the calculation, but round every intermediate result to four significant figures!
Problem 25 Prove Heron’s formula for the area of a triangle with sides $a$, $b$, and $c$, i.e. equation (5).

Problem 26 In Problem 25 we showed that Heron’s formula is correct, but in Problem 23 we discovered that round off errors can be disastrous. Here we investigate an alternative formulation of Heron’s formula. Show that

$$A^2 = \frac{1}{16}(a + b + c)(b + c - a)(a + c - b)(a + b - c) \quad (6)$$

Now return to the matrix given in Problem 23 and use the above formula to compute $A^2$. Remember to round each intermediate result to four significant figures.

3 Solution of quadratic equations

Problem 27 Let $A \neq 0$ and consider the familiar problem of solving the quadratic equation

$$Ax^2 + Bx + C = 0.$$

Derive the solution formula from scratch. Remember to justify every step! Show that the product of the two roots is equal to $C/A$ and that the sum of the two roots is the additive inverse of $B/A$.

Problem 28 Find the exact solution of the quadratic equation

$$x^2 - 1000x + 1 = 0.$$
using the standard formula. Repeat your calculations, but round each intermediate result to 4 significant figures. Which root is calculated with the smallest relative error. Use this root to compute an accurate value of the other root!

**Problem 29** Find the exact solutions of the quadratic equation

\[ x^2 + 2.011x + 1.011 = 0 \]

using the standard formula derived in Problem 27. Then repeat the calculations, but round each intermediate result to four significant figures.

**Problem 30** Find the exact solutions of the quadratic equation

\[ x^2 + x + 10^{-6} = 0 \]

using the standard formula derived in Problem 27. Then repeat the calculations, but round each intermediate result to four significant figures. Which root is calculated with the smallest relative error. Use this root to compute an accurate value of the other root!

**Problem 31** Consider the quadratic equation

\[ 10^{20}x^2 - 3 \cdot 10^{20}x + 2 \cdot 10^{20} = 0 \]

What are the exact roots? What happens if you try to solve this problem on a computer which uses single precision arithmetic? Can you overcome this problem by scaling the equation?

**Problem 32** Consider the quadratic equation

\[ 10^{-25}x^2 - 3x + 2 \cdot 10^{-25} = 0. \]

What are the exact roots? Can you solve this equation on a computer which uses single precision arithmetic? What happens if you scale this equation with, say, \(10^{-25}\)?

### 4 Mathematical induction

Mathematical induction is a standard technique which you must master as a scientist. The following problems focus on how to write a proof which uses mathematical induction using the correct language. You will derive some results which we will need in the course, but in the beginning the goal is to master the proper notation.

The principle of mathematical induction is really a statement about the nature of the integers. It is one of the *defining* properties of the set of positive integers \(\mathbb{N}\). Specifically, if \(V \subseteq \mathbb{N}\) satisfies

- \(1 \in V\)
• $a \in V \Rightarrow a + 1 \in V$

then $V = \mathbb{N}$. This sounds deceptively simple, but it is one of the truly powerful properties of $\mathbb{N}$.

**Problem 33** Let $a \in \mathbb{R}$ be given and consider the sequence $\{x_n\}_{n=1}^{\infty}$ given by

$$
x_0 = a,
\quad x_j = x_{j-1}^2, \quad j \in \mathbb{N}.
$$

Show that

$$\forall n \in \mathbb{N} : x_n = a^{2^n}.$$

**Problem 34** Let $a \in (0, \infty)$ be given and consider the sequence $\{x_n\}_{n=1}^{\infty}$ given by

$$
x_0 = a,
\quad x_j = \sqrt{x_{j-1}}, \quad j \in \mathbb{N}.
$$

Show that

$$\forall n \in \mathbb{N} : x_n = a^{2^{-n}}.$$

**Problem 35** Let $n \in \mathbb{N}$. Show that

$$\sum_{j=1}^{n} j = \frac{n(n+1)}{2}.$$

**Problem 36** Let $n \in \mathbb{N}$. Show that

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}.$$

**Problem 37** Let $n \in \mathbb{N}$. Show that

$$\sum_{j=1}^{n} j^3 = \left(\frac{n(n+1)}{2}\right)^2.$$

**Problem 38** Let $n \in \mathbb{N}$. Show that

$$\forall x, y \in \mathbb{R} : \quad (x + y)^n = \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}.$$
Problem 39 Let $x \in \mathbb{R}$ and let $p : \mathbb{R} \to \mathbb{R}$ be the polynomial

$$p(x) = \sum_{j=0}^{n} a_j x^j$$

and consider Algorithm 1.

**Algorithm 1** Horner’s method

1. $h_0 := a_n$
2. for $j = 1, \ldots, n$ do
3. \hspace{1em} $h_j := h_{j-1}x + a_{n-j}$
4. end for

Prove that $h_n = p(x)$ using mathematical induction.

**Hint** Define auxiliary polynomials $p_j$ given by

$$p_j(x) = \sum_{k=0}^{j} a_{n-k} x^{j-k}, \quad j \in \{0, 1, 2, \ldots, n\},$$

and define

$$V = \{ j \in \{0, 1, 2, \ldots, n\} : h_j = p_j(x) \}.$$

5 Miscellaneous problems

Problem 40 Let $n$ be a positive integer and let $1 \leq j \leq n$ be another integer. Show that

$$\binom{n}{j-1} + \binom{n}{j} = \binom{n+1}{j}.$$

**Remark 3** There is an easy way to remember this identity. Suppose you attend a class together with $n$ other students and $j$ students must give one presentation. From the teachers perspective there are $\binom{n+1}{j}$ ways of picking the $j$ students. However, what matters to you is whether you must give a talk or not! If you are to give a talk, then the remaining $j-1$ speakers must be picked from the other $n$ student and there are $\binom{n}{j-1}$ ways of doing this. If you are not to give a talk, then all the $j$ speakers must be chosen from the $n$ other students. This can be done in $\binom{n}{j}$ different ways. Obviously, the total count does not depend on who is counting, so we must have

$$\binom{n}{j-1} + \binom{n}{j} = \binom{n+1}{j}.$$
Problem 41 Compute the derivative of the function $a$ given by

$$a(j) = X x^j (y - j)^\beta$$

where $X$, $x$, $y$, and $\beta$ are all real numbers, with $x > 0$, and $y > 0$.

Problem 42 Prove or disprove the equality

$$\frac{1}{\sqrt{2} - 1} = \sqrt{2} + 1.$$ 

Problem 43 Is it true that

$$\sqrt{41} = \frac{156843854425524193}{24494894774743008}$$

Justify your answer!

6 Fundamental functions

Problem 44 Consider the function $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) = \int_1^x \frac{1}{t} dt.$$ 

1. Show that $f$ is strictly increasing, i.e.

   $$\forall x, y \in (0, \infty) : x < y \Rightarrow f(x) < f(y)$$

2. Show that

   $$f(x) \to \infty, \quad x \to \infty, \quad x \in (0, \infty).$$

   **Hint** Begin by showing

   $$\int_{2^j}^{2^{j+1}} \frac{1}{t} dt \geq \frac{1}{2}$$

   for $j \in \{0, 1, 2, \ldots\}$, and that

   $$f(2^n) \geq \frac{n}{2}.$$

   for $n \in \{1, 2, \ldots\}$.

3. Show that

   $$\forall a > 0 : f(a^{-1}) = -f(a)$$
4. Show that 
\[ f(x) \to -\infty, \quad x \to 0, \quad x \in (0, \infty). \]

5. Show that 
\[ \forall a, b \in (0, \infty) : f(ab) = f(a) + f(b). \]

**Remark 4** The natural logarithm is *defined* by
\[ \log(x) = \int_1^x \frac{1}{t} \, dt \]
You have just derived some of the fundamental properties of this function entirely from scratch.

**Problem 45** Reconsider the function \( f \) defined in Problem 44.

1. Show that \( f \) has an inverse function \( g : \mathbb{R} \to (0, \infty) \)

2. Show that \( g \) is differentiable and 
\[ \forall y \in \mathbb{R} : \quad g'(y) = g(y) \]
and \( g(0) = 1 \).

3. Show that 
\[ \forall x, y \in \mathbb{R} : \quad g(x + y) = g(x)g(y). \]

**Remark 5** The natural exponential function \( \exp \) is defined as the inverse function of the natural logarithm. You have just proven from scratch that
\[ e^{x+y} = e^x e^y \]
The differential equation satisfied by \( \exp \) can be used to prove that
\[ e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}, \]
an expression which is frequently taken as the definition of \( \exp \).

**Problem 46** A plane rotation can be represented using the matrix
\[ A(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \]
where \( \theta \) is the angle measured in radians.

1. Prove without doing any calculations at all that
\[ \forall \theta, \psi \in \mathbb{R} : A(\theta)A(\psi) = A(\theta + \psi). \]
2. Show that

\[
\cos(\theta + \psi) = \cos \theta \cos \psi - \sin \theta \sin \psi, \quad (9)
\]
\[
\sin(\theta + \psi) = \sin \theta \cos \psi + \cos \theta \sin \psi. \quad (10)
\]

**Problem 47** Show that

\[
\tan(\theta + \psi) = \frac{\tan \theta + \tan \psi}{1 - \tan \theta \tan \psi}
\]

provided \(\cos \theta, \cos \psi\) and \(\cos(\theta + \psi)\) are all nonzero.

**Problem 48** Show that

\[
\forall \theta \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right) : \tan 2\theta = \frac{2 \tan \theta}{1 - \tan^2 \theta}.
\]

**Problem 49** Show that

\[
\forall \theta \in \left(-\frac{\pi}{8}, \frac{\pi}{8}\right) : \tan 4\theta = \frac{4 \tan \theta(1 - \tan^2 \theta)}{(1 - \tan^2 \theta)^2 - 4 \tan^2 \theta}
\]

**Problem 50** Show that

\[
\frac{\pi}{4} = 4 \arctan \left(\frac{1}{5}\right) - \arctan \left(\frac{1}{239}\right). \quad (11)
\]

### 7 Infinite series

**Problem 51** Consider the infinite series

\[
s = \sum_{n=1}^{\infty} \frac{1}{n^2}.
\]

1. Show that the series is convergent with \(s \geq 1\) and that

\[
\sum_{j=N+1}^{\infty} \frac{1}{j^2} \leq \int_{N}^{\infty} \frac{1}{x^2} \, dx = \frac{1}{3N^3}
\]

2. How many terms do you need to include in order to compute \(s\) with a relative error less than, say, \(\tau = 10^{-15}\) assuming exact arithmetic.
3. Estimate the rounding error committed by adding such a large number of floating point numbers? Can you guarantee that the error is less than \( \tau = 10^{-15} \).

**Problem 52** Consider the infinite series

\[
s = \sum_{j=1}^{\infty} \frac{1}{j^4}.
\]

1. Show that the series is convergent with \( s > 1 \) and that

\[
\sum_{N+1}^{\infty} \frac{1}{j^4} \leq \int_N^{\infty} \frac{1}{x^4} dx = \frac{1}{5N^5}
\]

2. How many terms do you need to include in order to compute \( s \) with a relative error less than, say, \( \tau = 10^{-15} \) assuming exact arithmatic. (Compare with Problem 51).

3. Estimate the rounding error induced by adding such a large number of terms. Can you guarantee that the rounding error is less than \( \tau = 10^{-15} \)?

**Remark 6** It is possible to prove using Fourier series methods that

\[
\pi^2 \frac{6}{\pi} = \sum_{n=1}^{\infty} \frac{1}{n^2} \quad \text{and} \quad \pi^4 \frac{90}{\pi^4} = \sum_{n=1}^{\infty} \frac{1}{n^4}
\]

and we can certainly use these identities to compute approximations of \( \pi \). However, as we shall see later, there are more efficient algorithms for computing \( \pi \).

**Problem 53** Show that

\[
\forall x \neq 1 : \sum_{j=0}^{n} x^j = \frac{1-x^{n+1}}{1-x}
\]

and if \( |x| < 1 \), then the infinite sequence

\[
s = \sum_{j=0}^{\infty} x^j
\]

is convergent with sum \( s = \frac{1}{1-x} \).

**Problem 54** The purpose of this problem is to develop techniques powerful enough to derive the identity

\[
\pi \frac{4}{4} = \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1}
\]  \( (12) \)
1. Let \( n \in \mathbb{N} \). Show that 
\[
\forall x \in \mathbb{R} : \frac{1}{1 + x^2} = \sum_{j=0}^{n} (-1)^j x^{2j} + R_n(x)
\]
where the remainder term is given by 
\[
R_n(x) = (-1)^{n+1} \frac{x^{2n+2}}{1 + x^2}
\]

2. Show that 
\[
\pi = \int_0^1 \frac{1}{1 + x^2} dx = \sum_{j=0}^{n} \frac{(-1)^j}{2n+1} + \int_0^1 R_n(x) dx
\]
for all \( n \in \mathbb{N} \).

3. Show that 
\[
\left| \int_0^1 R_n(x) dx \right| \leq \int_0^1 x^{2n+2} dx = \frac{1}{2n+3}
\]
for all \( n \in \mathbb{N} \).

4. Show that the infinite sequence 
\[
s = \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1}
\]
is convergent with sum \( s = \frac{\pi}{4} \).

5. How many terms do you need to include in order to approximate \( \frac{\pi}{4} \) with an error which is less than \( \frac{1}{100} \) in absolute value?

### 8 Linear approximation of real functions

**Problem 55** Consider the real functions \( f \) and \( g \) given by 
\[
f(x) = \sqrt{x}, \quad \text{and} \quad g(x) = \frac{1}{3} x + \frac{17}{24}
\]
Show that 
\[
\forall x \in [1, 4] : -\frac{1}{24} \leq f(x) - g(x) \leq \frac{1}{24}.
\]

**Problem 56** Consider the real functions \( f \) and \( g \) given by 
\[
f(x) = \sqrt{x}, \quad g(x) = ax + b
\]
where

\[ a = \frac{1}{7}, \quad \text{and} \quad b = \frac{3}{7} \sqrt[3]{\frac{7}{3}} + \frac{1}{3} \sqrt[3]{\frac{7}{3}} \]

Show that

\[ \forall x \in [1, 8] : |f(x) - g(x)| \leq c = \frac{3}{7} - \frac{1}{3} \sqrt[3]{\frac{7}{3}} < 8.061 \times 10^{-2}. \]

**Remark 7** We observe that it is very inexpensive to compute a good approximation of \( \sqrt{x} \) for \( x \in [1, 8] \). Such an initial guess can be refined using, say, Newton’s method. A plot of \( f \), \( g \) and the function \( \psi(x) = \phi(x)/c \), is given in Figure 2.

**Problem 57** Consider the real functions \( f \) and \( g \) given by

\[ f(x) = \arcsin(x), \quad \text{and} \quad g(x) = ax + b, \]

where \( x \in [0, 1] \) and

\[ a = \frac{\pi}{2}, \quad \text{and} \quad b = \frac{1}{2} \arcsin \left( \frac{\sqrt{a^2 - 1}}{a} \right) - \frac{1}{2} \sqrt{a^2 - 1} \]

Show that

\[ \forall x \in [0, 1] : |f(x) - g(x)| \leq |b| < 1.653 \times 10^{-1}. \]
9 Taylor polynomials and series

Problem 58 Find the Taylor polynomial of order $n$ for the function

$$f(x) = \exp(x)$$

at the point $x_0 = 0$. Estimate the size of corresponding error term.

Problem 59 Find the Taylor polynomial of order $n = 2m + 1$ for the function

$$f(x) = \sin(x)$$

at the point $x_0 = 0$. Estimate the size of corresponding error term.

Problem 60 Find the Taylor polynomial of order $n = 2m$ for the function

$$f(x) = \cos(x)$$

at the point $x_0 = 0$.

Problem 61 Consider the real function $f$ given by

$$f(x) = \begin{cases} 
\exp\left(-\frac{1}{1-x^2}\right) & -1 < x < 1 \\
0 & |x| \geq 1
\end{cases}$$

Show that $f$ is infinitely often differentiable and show that the Taylor polynomial of degree $n$ for $f$ at the point $x_0 = 1$ is identically zero.

Problem 62 Find the $n$ order Taylor polynomial for the function

$$f(x) = \cosh(x) = \frac{e^x + e^{-x}}{2}$$

at the point $x_0 = 0$. Estimate the value of the corresponding error term.

Problem 63 Consider the problem of computing

$$f(x) = e^x,$$

for $x \in [1, 2]$.

1. Find the $n$ order Taylor polynomial $p_n$ for $f$ at the point $x_0 = 0$.

2. Show that the absolute error satisfies

$$|f(x) - p_n(x)| \leq e^2 \cdot \frac{2^{N+1}}{(N + 1)!}$$

for all $x \in [1, 2)$. 

3. Show that the relative error satisfies
\[
\frac{|f(x) - y_n(x)|}{|f(x)|} \leq e \cdot \frac{2^{N+1}}{(N+1)!}
\]
for \( x \in [1, 2) \).

4. Explain why there is no point at all in choosing \( N = 50 \) on a machine which use IEEE double precision floating point arithmetic. What is the largest number of terms it makes sense to use?

**Problem 64** Consider the problem of computing
\[
f(x) = \sin x = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} x^{2j+1}
\]
for \(|x| < \delta\) where \( 0 < \delta < 1 \).

1. Show that the 4th order Taylor polynomial for \( f \) at the point \( x_0 = 0 \) is given by
\[
g(x) = x - \frac{1}{6} x^3
\]
2. Show that
\[
0 < g(x) < f(x) < g(x) + \frac{1}{120} x^5 \tag{13}
\]
for all \( x \in (0, \delta) \).

**Hint** Show that series for \( f(x) = \sin(x) \) is alternating and the terms decay monotonically when \( 0 \leq x < \sqrt{6} \).

3. Show that the relative error satisfies
\[
\frac{|f(x) - g(x)|}{|f(x)|} \leq \frac{1}{120} \cdot \frac{x^4}{1 - \frac{1}{6} x^2} \leq \frac{\delta^4}{100} \tag{14}
\]
for all \( x \in (0, \delta) \).

4. What is the largest value of \( \delta \) for which \( g(x) \) can be used to approximate \( f(x) \) with a relative error of at most \( \tau = 10^{-16} \) for all \( x \in (0, \delta) \)?

**Problem 65** Consider the function \( f : [-4, \infty) \to \mathbb{R} \) given by
\[
f(x) = \sqrt{4 + x} - 2
\]
Find a numerically reliable way to compute \( f(x) \) for all \( x \). Why is there even a problem for small values of \(|x|\)?
**Problem 66** Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \frac{\sin(x) - x}{x^3}$$

for $x \neq 0$. Which value should you assign to $f(0)$ in order to have a continuous function? Find a numerically reliable way to compute $f(x)$ for $x \approx 0$. Why is there even a problem for small values of $|x|$?

**Problem 67** Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \sinh(x) = \frac{\exp(x) - \exp(-x)}{2}.$$

Find a numerically reliable way to compute $f(x)$ for all $x$. Where do you need to be careful?

**Problem 68** Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \frac{1 - \cos(x)}{x^2}$$

for $x \neq 0$. Which value should you assign to $f(0)$ in order to have a continuous function? Find a numerically reliable way to compute $f(x)$ for $x \approx 0$. Why is there even a problem for small values of $|x|$?

**Problem 69** Consider the real valued function $f$ given very loosely by the assignment

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

Where is this function even defined? Find a numerically reliable way to evaluate $f(x)$ for all relevant $x$.

### 10 Non-linear equations

**Problem 70** Find all real solutions of the equation

$$\sqrt{x + 3} = \sqrt{1 - x} + \sqrt{1 + x}$$

(17)

**Problem 71** Find all real solutions of the equation

$$\sqrt{x + 3} = \sqrt{1 - x} - \sqrt{1 + x}.$$  

(18)

**Problem 72** Find all real solution of the equation

$$\sqrt{x + 3} = \sqrt{1 + x} - \sqrt{1 - x}$$

(19)
11 Numerical Integration

**Problem 73** Let \( f : [0, 1] \rightarrow \mathbb{R} \) be given by
\[
f(x) = x^3
\]
Apply the composite trapezoidal rule to the problem of calculating \( I = \int_0^1 f(x)\,dx \) and compare the approximation with the exact result.

**Problem 74** Let \( f : [0, 1] \rightarrow \mathbb{R} \) be given by
\[
f(x) = e^x
\]
and consider the integral
\[
I = \int_0^1 f(x)\,dx.
\]
Let \( N > 0 \) be an integer, and let \( T_h(f) \) be the trapezoidal sum corresponding to the uniform stepsize \( h = 1/N \). Show that
\[
T_h(f) = \frac{1}{2}(e-1) \frac{h}{\tanh(h/2)}, \quad \text{where} \quad \tanh(x) = \frac{\sinh(x)}{\cosh(x)},
\]
and show by an explicit calculation that
\[
T_h(f) \to \int_0^1 f(x)\,dx, \quad N \to \infty, \quad N \in \mathbb{N}, \quad Nh = 1.
\]
Moreover, show that
\[
T_h(f) = (e - 1) \left[ 1 + \frac{h^2}{12} - \frac{h^4}{720} + \frac{h^6}{30240} + O(h^8) \right].
\]
You are welcome to use a program like Mathematica as the series expansion is a bit tedious to derive.

**Problem 75** Let \( f : [0, 1] \rightarrow \mathbb{R} \) be given by
\[
f(x) = x^3
\]
Compute the composite Simpson sum for this integral corresponding to a uniform stepsize of \( h \) and compare the approximation with the exact result.

**Problem 76** Let \( f : [0, 1] \rightarrow \mathbb{R} \) be given by
\[
f(x) = e^x
\]
Apply the composite Simpson rule to the problem of calculating \( \int_0^1 f(x)\,dx \) and compare the approximation with the exact result.
Problem 77 Show that improper integral

\[ I = \int_0^\infty xe^{-x} \, dx \]

exists and has the value \( I = 1 \). Show that the trapezoidal sum \( T_h \) corresponding to a uniform stepsize \( h \) is

\[ T_h = e^h \left( \frac{h}{e^h - 1} \right)^2 \]

Moreover, show that

\[ I - T_h(f) = \frac{1}{12} h^2 - \frac{1}{240} h^4 + \frac{1}{6048} h^6 + O(h^8) \]

12 Numerical solution of differential equations

Problem 78 Let \( a : \mathbb{R} \to \mathbb{R} \) be a continuous function and define

\[ A(t) = \int_{t_0}^t a(s) \, ds. \]

Show that the initial value problem

\[ y'(t) = a(t)y(t), \quad y(t_0) = y_0 \]

has a unique solution given by

\[ y(t) = y_0 \exp(A(t)) \quad (20) \]

Problem 79 Find the general solution of the differential equation

\[ y'(t) = ay(t) + b, \quad (21) \]

where \( a \neq 0 \) and \( b \) are real numbers.