Problem 1 Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

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

1. (5 points) Show that f is differentiable and strictly increasing for all $x \in \mathbb{R}$.

   Solution The function $g : \mathbb{R} \to (0, \infty)$ given by

   $$g(x) = e^x$$

   is a well known differentiable function. Moreover, since

   $$f(x) = \frac{1}{2}\left(g(x) + 1/g(x)\right)$$

   is build from g using one division, one addition and one scalar multiplication, f is differentiable. We have

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

   from which it follows that f is strictly increasing.

2. (5 points) The following MATLAB commands have been used to generate the plot in Figure 1.

```
>> k=21;
>> x=single(linspace(-1,1,129)*2^-k);
>> f=@(x)(exp(x)-exp(-x))/2;
>> plot(x,f(x))
```

   Explain why this is clearly not an accurate representation of the graph of f on the interval $[-2^{-21}, 2^{-21}]$?

   Solution This is not the plot of a function which is strictly increasing! It appears that there are many solutions of the equation $f'(x) = 0$, which is direct violation of the fact that $f'(x) > 0$ for all x.

3. (5 points) Consider the nominator of $f(x)$, i.e. the expression

   $$N(x) = e^x - e^{-x}.$$

   Show that we do not have to worry about catastrophic cancellation when $x > \frac{\log(2)}{2}$.

   Solution In general, we do not have to worry about catastropic cancellation in a subtraction $a - b$, if, say, $a > 2b > 0$. In our case, we consider the subtraction $d(x) = a(x) - b(x)$ where $a(x) = e^x$ and $b(x) = e^{-x}$. We have

   $$a(x) > 2b(x) \Leftrightarrow e^{2x} > 2 \Leftrightarrow 2x > \log(2),$$

   or equivalently $x > \frac{\log(2)}{2}$.

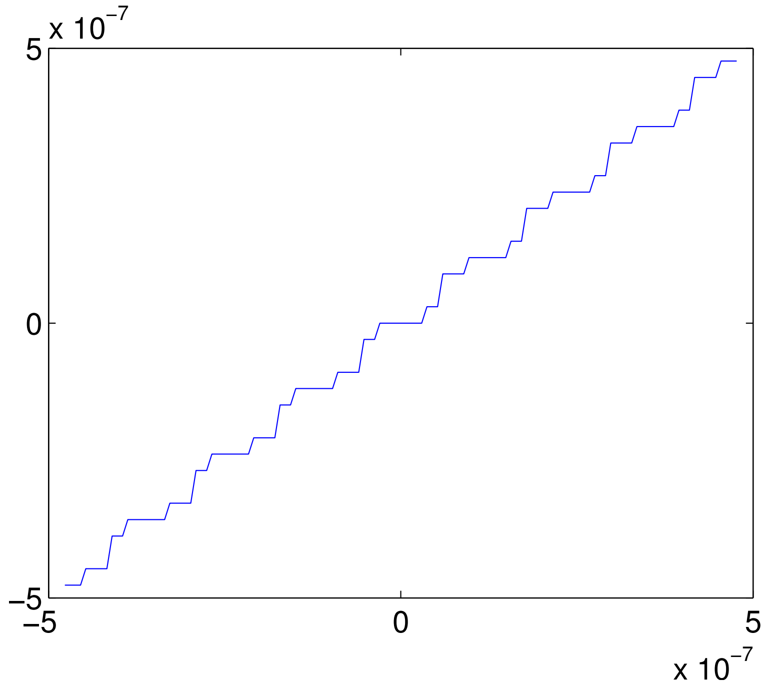

Figure 1: An untrustworthy plot of the function f on an interval around zero.

4. (2 points) Let p_n be the Taylor polynomial for f of order n at the point $x_0 = 0$. Show that

$$p_7(x) = x + \frac{x^3}{6} + \frac{x^5}{120} + \frac{x^7}{5040}.$$

Solution It is well known that

$$g(x) = e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$

Therefore, the Taylor series for f is merely

$$f(x) = \frac{e^x - e^{-x}}{2} = \sum_{j=0}^{\infty} \frac{x^{2j+1}}{(2j+1)!},$$

The Taylor polynomial of order 7 is obtained by truncating the series immediately after the term which contains x^7, i.e. $j = 3$.

5. (6 points) Show that

$$\frac{|f(x) - p_7(x)|}{|f(x)|} \le \frac{|x|^8}{8!}, \quad x > 0.$$

Solution Let $x > 0$. By Taylor's formula, there exists $\xi \in (0, x)$ such that

$$f(x) - p_7(x) = \frac{f^{(8)}(\xi)}{8!} x^8$$

Now, $f(x) = \sinh(x)$, so $f'(x) = \cosh(x)$ and $f''(x) = \sinh(x)$. It follows that

$$f^{(8)}(x) = f(x)$$

and

$$\frac{f(x) - p_7(x)}{f(x)} = \frac{f^{(8)}(\xi)}{f(x)} \frac{x^8}{8!}.$$

Since f is monotone increasing for all x, and positive for all $x > 0$, we have

$$\left| \frac{f(x) - p_7(x)}{f(x)} \right| = \left| \frac{f^{(8)}(\xi)}{f(x)} \frac{x^8}{8!} \right| = \frac{f^{(8)}(\xi)}{f(x)} \frac{x^8}{8!} \leq \frac{x^8}{8!} = \frac{|x|^8}{8!}.$$

6. (2 points) Show that p_7 approximates f with a relative error τ which is smaller than the single precision round-off error, i.e. $u = 2^{-24}$, on the interval $(0, \log(2)/2]$.

Solution We have already have

$$\left| \frac{f(x) - p_7(x)}{f(x)} \right| \leq \frac{|x|^8}{8!}$$

for all $x > 0$. In particular,

$$\left| \frac{f(x) - p_7(x)}{f(x)} \right| \leq \frac{1}{8!} \left(\frac{\log(2)}{2} \right)^8 \approx 5.162299 \times 10^{-9} < 2^{-24}$$

Problem 2 An infinitely differentiable function $f : [0,1] \to \mathbb{R}$ has been integrated numerically on the interval $[0,1]$ using the standard trapezoidal rule $T = T(h)$ and stepsizes $h = h(k) = 2^{-k}$ for many different values of k. The results and some auxiliary calculations are given in the table below.

k	Th	(Th-T2h)/3	(T2h-T4h)/(Th-T2h)
24	-0.1056405560	-5.2828e-15	0.0674255692
23	-0.1056405560	-3.5620e-16	35.2987012987
22	-0.1056405560	-1.2573e-14	2.6405445180
21	-0.1056405560	-3.3200e-14	4.6455343458
20	-0.1056405560	-1.5423e-13	3.8538436160
19	-0.1056405560	-5.9439e-13	4.0087321291
18	-0.1056405560	-2.3828e-12	4.0049681024
17	-0.1056405560	-9.5428e-12	3.9991143549
16	-0.1056405559	-3.8163e-11	4.0001060634
15	-0.1056405558	-1.5266e-10	3.9999743333
14	-0.1056405554	-6.1062e-10	4.0000016515
13	-0.1056405535	-2.4425e-09	4.0000000947
12	-0.1056405462	-9.7699e-09	3.9999994938
11	-0.1056405169	-3.9080e-08	3.9999985076
10	-0.1056403996	-1.5632e-07	3.9999939668
9	-0.1056399307	-6.2527e-07	3.9999758652
8	-0.1056380549	-2.5011e-06	3.9999034071
7	-0.1056305516	-1.0004e-05	3.9996127750
6	-0.1056005394	-4.0012e-05	3.9984373960
5	-0.1054805023	-1.5999e-04	3.9935271855
4	-0.1050005413	-6.3891e-04	3.9703436822
3	-0.1030838038	-2.5367e-03	3.8067034455
2	-0.0954736974	-9.6565e-03	1.4391333304
1	-0.0665042792	-1.3897e-02	
0	-0.0248134239		

1. (5 points) Explain, why it is immediately clear that computed values of the tell-tale fraction
$$\frac{T_{2h} - T_{4h}}{T_h - T_{2h}}$$
are completely wrong for $k > 19$.

   Solution In exact arithmetic, the fractions should tend to 4. However, for $k > 19$ is is crystal clear that computed fractions are no longer displaying this behavior.

2. (4 points) Explain, why the expression
$$T_h - T_{2h}$$

cancelled catastrophically for large values of k.

Solution The real numbers T_h and T_{2h} satisfy

$$T_h, T_{2h} \to \int_0^1 f(x)dx, \quad h \to 0, \quad h > 0,$$

simply because f is two times differentiable with a continuous second derivative. It follows that T_h and T_{2h} are very close, when h is very small. As a result, the expression $T_h - T_{2h}$ will suffer from catastropic cancellation.

3. (5 points) Explain why the computed approximations of the integral are inaccurate for small values of k.

   Solution Small values of k correspond to large values of the stepsize h. The trapezoidal sum is formed by approximating f with a linear function on each subinterval of length h. In general, this is not a good approximation, unless h is small.

4. (7 points) Determine the range of k where you are confident that you can trust the error estimates. Remember to justify your choice!

   Solution If the fractions are computed in exact arithmetic, then they will converge to 4 monotonically, either from above 4 or from below 4. In practice we see deviation from this behavior because of roundoff errors in the computation of the sum T_h. The computed value of our fraction is approaching 4 from below for $k = 3, 4, 5, \dots, 12$. At $k = 13$, the value has jumped to the other side of 4, something which should not happen in exact arithmetic. I would trust the sign, magnitude and the first couple of digits of the error estimates for $k = 4, 5, \dots, 12$ simply because the fraction is not only close to 4, but converging monotonically to 4 as h is decreased further. In all likelyhood, the error estimate is still good at $k = 13$, but further investigation is required to determine that.

5. (5 points) Determine the smallest value of k from which the value of the integral can be approximated with a *relative* error which is smaller than $\tau = 10^{-7}$.

   Solution We are handed the correct estimats of the error, i.e. column 3. By inspection we see that the absolute value of integral is slightly larger than $\frac{1}{10}$. Therefore, multiplying the error estimates with a factor of 10 will give us an estimate of the relative error. We observe that the absolute value of the relative error is less than $9.7699 \times 10^{-8} < \tau$ for $k = 12$. Similarly, the absolute value of relative error is abot 3.9080×10^{-7}, i.e. too large for $k = 11$. Since $k = 12$ falls in the range were we trust the error estimate, we conclude that $k = 12$ is the smallest integer where the relative error is bounded by $\tau = 10^{-7}$.

Problem 3 Let $y \in (0,1)$ and consider the non-linear equation

$$g(y) = 0$$

where

$$g(y) = \frac{\sqrt{1-y^2}}{y} - \tan(y).$$

1. (5 points) Show that this equation has at least one solution on the interval $(0,1)$.

   Solution We have

   $$g(y) \to \infty, \quad y \to 0, \quad y \in (0,1)$$

   and

   $$g(y) \to -\tan(1), \quad y \to 1, \quad y \in (0,1)$$

   In short, g assumes both (strictly) positive and (strictly) negative values on the interval $(0,1)$. Since g is continuous, there must be at least one point $\xi \in (0,1)$ where $g(\xi) = 0$.

2. (5 points) Show that the solution is unique.

   Solution We have

   $$\begin{aligned} g'(y) &= -\frac{\sqrt{1-y^2} - \frac{y(-2y)}{2\sqrt{1-y^2}}}{y^2} - (1 + \tan^2(y) \\ &= -\sqrt{1-y^2}\left(\frac{1 + \frac{y^2}{(1-y^2)}}{y^2}\right) - 1 - \tan^2(y) < 0. \end{aligned}$$

   It follows, that g is strictly decreasing and there can be only one solution of the equation $g(x) = 0$.

3. (7 points) Write down an iteration which is certain to converge to the solution, provided you begin with a good initial guess.

   Solution If we have a good initial guess x_0, then Newton's method is a good candidate. We have

   $$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

4. (8 points) The following table displays the results of applying Newton's method to the problem at hand.

```
n            x(n)                           g(x(n))
--------------------------------------------------------
0      5.000000000000000e-01      1.185748317725087e+00
1      7.003884584051097e-01      1.761416298335665e-01
2      7.389598589869100e-01      5.697906526650476e-04
3      7.390851339148572e-01     -3.182480501351392e-09
4      7.390851332151607e-01     -1.110223024625157e-16
5      7.390851332151607e-01     -1.110223024625157e-16
6      7.390851332151607e-01     -1.110223024625157e-16
7      7.390851332151607e-01     -1.110223024625157e-16
8      7.390851332151607e-01     -1.110223024625157e-16
9      7.390851332151607e-01     -1.110223024625157e-16
```

In exact arithmetic, Newton's iteration converges quadratically, and the number of correct digits should double for every iteration. Explain why the computed numbers stagnate after $n = 4$.

Solution The best we can hope for is to obtain the floating point representation of the root ξ, i.e. a relative error which is bounded by u. Iterations beyond the point $|x_n - \xi|/|\xi| < u$ are pointless, because we will never see the numbers x_n. At best we can obtain the floating point representation of x_n. In floating point arithmetic, Newton's iteration will either lock on a particular floating number, as in our case, or cycle through a small set of numbers in the immediate vicinity of the root.

Problem 4 The "rage" virus has escaped from the laboratory at the heart of the green zone in London and the zombies are attacking the civilian population. Table 1 gives the number of infected during the initial phase.

t (minutes)	infected
0	1
2	6
4	19
6	40
8	69

Table 1: The number of zombies as function of time during the first few minutes after the outbreak.

1. (8 points) Find a polynomial of degree at most 2 which fits the initial data.

   Solution Let $z(t)$ denote the number of zombies at time t. We use the first three nodes, i.e. $t = 0, 2, 4$ to obtain a polynomial of the form

   $$p(x) = c_0 + c_1 t + c_2 t(t-2)$$

   which matches z at these nodes. We have

   $$z(0) = 1 = p(0) = c_0.$$

   Similarly, we have

   $$z(2) = 6 = p(2) = c_0 + 2c_1 = 1 + 2c_1 \Rightarrow c_1 = \frac{5}{2}$$

   and finaly we have

   $$z(4) = 19 = p(3) = c_0 + 4c_1 + 8c_2 = 1 + 10 + 8c_2 \Rightarrow c_2 = 1$$

   In summary,

   $$p(t) = t(t-2) + \frac{5}{2}t + 1$$

   is polynomial of order 2 which interpolates f at the nodes $t = 0, 2, 4$. We also have

   $$p(6) = 40 = z(6), \quad p(8) = 69$$

   so the polynomial appears to describe the zombie populations very well.

2. (7 points) Estimate the rate of infection, i.e. new zombies/minute at $t = 8$ minutes.

   Solution The rate of infection is $z'(t)$. We can only differentiate p. We have

   $$p'(t) = (t-2) + t + \frac{5}{2} = 2t + \frac{1}{2}$$

3. (10 points) The garrison has the capacity to kill 50 zombies/minute using conventional small arms. Assuming that our model continues to hold, at which time will it be impossible to stabilize the zombie population at a fixed number of individuals.

   Solution The situation get out of control if $z'(t) > 50$. The best we can do is to solve $p'(t) > 50$. We have

$$p'(t) > 50 \Leftrightarrow 2t + \frac{1}{2} > 50 \Leftrightarrow t > 24.75$$

   In short, the situation requires unconventional weapons after about 25 minutes.