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

$$f(x) = \cosh(x) - \cos(x).$$

1. (5 points) The following `MATLAB` commands have been issued

```
>>f=@(x)cosh(x)-cos(x);
>>x=linspace(-1,1,1024)*2^(-10); x=single(x);
>>plot(x,f(x))
```

and the result is given as Figure 1. List as many differences between this plot and the correct graph of f as you can.

Solution The function f is differentiable and $f'(x) = \sinh(x) + \sin(x)$ satisfies that $f'(x) > 0$ for $x > 0$ and $f'(x) < 0$ for $x > 0$. However, the plot does not show a function which is strictly increasing for $x > 0$ and strictly decreasing for $x < 0$. In fact, there plot gives the wrongful impression that the equation $f'(x) = 0$ has infinitely many solutions. Moreover, the sudden jumps from one value to another are inconsistent with a function f for which f' varies continuously around the value 0.

2. (10 points) Prove that the naive expression for f used in the previous question cannot cancel catastrophically when

$$x > \cosh^{-1}(2) = \log(2 + \sqrt{3}).$$

Solution If $x > 2y$, then catastrophic cancellation is not an issue, when we attempt to compute $d = x - y$. In our case we find that

$$x > \cosh^{-1}(2) \Rightarrow \cosh(x) > 2 \Rightarrow \cosh(x) > 2\cos(x)$$

which allows us to conclude the original expression for f can not cancel catastrophically for $x > \cosh^{-1}(2)$.

3. (10 points) Find a polynomial p for which you are certain the following two properties are true
   - For all x we have $f(x) - p(x) = O(x^{10})$.
   - Catastrophic cancellation cannot occur when evaluating p for small values of x.

Solution We try a Taylor expansion of f at the point $x_0 = 0$. We have

$$\cosh(x) = \sum_{j=0}^{\infty} \frac{x^{2j}}{(2j)!} = 1 + \frac{1}{2}x^2 + \frac{1}{4!}x^4 + \frac{1}{6!}x^6 + \frac{1}{8!}x^8 + O(x^{10}) \qquad (1)$$

$$\cos(x) = \sum_{j=0}^{\infty} (-1)^j \frac{x^{2j}}{(2j)!} = 1 - \frac{1}{2}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \frac{1}{8!}x^8 + O(x^{10}) \qquad (2)$$

Therefore

$$\cosh(x) - \cos(x) = 2\left(\frac{1}{2}x^2 + \frac{1}{6!}x^6\right) + O(x^{10}),$$

and

$$p(x) = 2\left(\frac{1}{2}x^2 + \frac{1}{6!}x^6\right).$$

satisfies the first property that we are looking for. Moreover, since the evaluation of

$$p(x) = x^2 + \frac{1}{180}(x^2)^3$$

can be done without any subtractions at all, there is no risk of catastrophic cancellation.

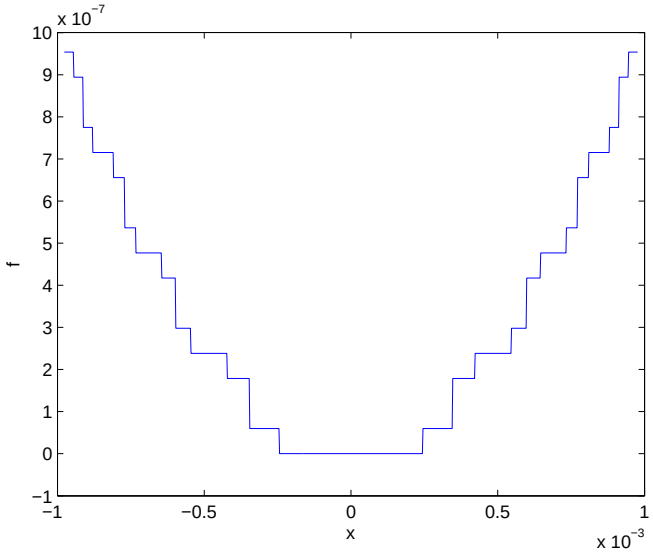

Figure 1: A plot of f as a function of x using a bad `MATLAB` implementation.

Problem 2 Consider the function $g : [-1, 1] \to \mathbb{R}$ given by

$$g(x) = (1 - |x|^3)^{\frac{1}{3}}.$$

1. (15 points) Simpson's rule has been used to generate many different approximations of the two integrals

$$I_1 = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(x)dx \quad \text{and} \quad I_2 = \int_{-1}^{1} g(x)dx$$

and all the results are given in Figure 2 and Figure 3. Unfortunately, we have forgotten which figure belongs to which integral!

N	Sh	(Sh-S2h)	(S2h-S4h)/(Sh-S2h)
131072	9.89322425564950e-01	-9.4369e-15	-6.117647e-01
65536	9.89322425564960e-01	5.7732e-15	9.230769e-01
32768	9.89322425564954e-01	5.3291e-15	8.958333e-01
16384	9.89322425564949e-01	4.7740e-15	-1.255814e+00
8192	9.89322425564944e-01	-5.9952e-15	-3.703704e-02
4096	9.89322425564950e-01	2.2204e-16	9.000000e+00
2048	9.89322425564950e-01	1.9984e-15	1.083333e+01
1024	9.89322425564948e-01	2.1649e-14	1.598974e+01
512	9.89322425564926e-01	3.4617e-13	1.599487e+01
256	9.89322425564580e-01	5.5369e-12	1.599390e+01
128	9.89322425559043e-01	8.8557e-11	1.597706e+01
64	9.89322425470486e-01	1.4149e-09	1.590904e+01
32	9.89322424055610e-01	2.2509e-08	1.564736e+01
16	9.89322401546288e-01	3.5221e-07	1.472995e+01
8	9.89322049334740e-01	5.1881e-06	1.215511e+01
4	9.89316861277389e-01	6.3061e-05	5.970799e+01
2	9.89253799880431e-01	3.7653e-03	0.000000e+00
1	9.85488530462065e-01	0.0000e+00	0.000000e+00

Figure 2: The first set of calculations, $2Nh$ equals the length of the relevant interval.

Examine the numbers carefully and give as many reasons as you can as to why Figure 3 contains the results obtained by applying Simpson's rule to the integral I_2.

Hint There are at least two reasons which are very different in nature.

Solution Reason 1 We are integrating the function g over to different intervals I_1 and I_2. Since $g \geq 0$ and $I_1 \subset I_2$ we must have

$$\int_{I_1} g(x)dx \leq \int_{I_2} g(x)dx$$

N	Sh	(Sh-S2h)	(S2h-S4h)/(Sh-S2h)
131072	1.76663869100275e+00	9.0100e-08	2.519844e+00
65536	1.76663860090236e+00	2.2704e-07	2.519845e+00
32768	1.76663837386342e+00	5.7210e-07	2.519849e+00
16384	1.76663780176041e+00	1.4416e-06	2.519856e+00
8192	1.76663636014718e+00	3.6327e-06	2.519870e+00
4096	1.76663272748955e+00	9.1538e-06	2.519897e+00
2048	1.76662357366548e+00	2.3067e-05	2.519953e+00
1024	1.76660050696801e+00	5.8127e-05	2.520063e+00
512	1.76654237998170e+00	1.4648e-04	2.520284e+00
256	1.76639589630144e+00	3.6918e-04	2.520724e+00
128	1.76602671582609e+00	9.3060e-04	2.521593e+00
64	1.76509611380473e+00	2.3466e-03	2.523265e+00
32	1.76274951440854e+00	5.9211e-03	2.526195e+00
16	1.75682842297277e+00	1.4958e-02	2.529481e+00
8	1.74187058957818e+00	3.7836e-02	2.521815e+00
4	1.70403504038579e+00	9.5414e-02	2.885182e+00
2	1.60862078851493e+00	2.7529e-01	0.000000e+00
1	1.33333333333333e+00	0.0000e+00	0.000000e+00

Figure 3: The second set of calculations, $2Nh$ equals the length of the relevant interval.

reflected in the tables. From Figure 2 we deduce that the value is of the appropriate integral is certainly less than 1.0. Similarly, Figure 3 relates to an integral which is certainly larger than 1.7. Therefore we conclude that Figure 3 relates to $\int_{I_2} g(x)dx$. **Reason 2** This perception is only strengthened when we examine the function and the tell tale fractions in detail. In Figure 2 we see fractions which converge monotonically towards 16 until rounding errors become a problem. This is the behavior which we expect from a function which is (sufficiently) differentiable. In Figure 3 we see fractions which converge monotonically towards $2^{\frac{4}{3}}$, i.e. not 2^4. This is the behavior we have seen from functions which are not differentiable on the entire interval of integration. It is clear that g is not differentiable at $x = \pm 1$, so g is not differentiable on I_2. Is g differentiable on I_1? There is only one problematic point, namely $x = 0$. However, the auxiliary function $\phi(x) = |x|^3 = |x|x^2$ is certainly differentiable at $x = 0$ because

$$\frac{\phi(x) - \phi(0)}{x - 0} = \frac{|x|x^2 - 0}{x - 0} = |x|x \to 0, \quad x \to 0, \quad x \neq 0.$$

The chain rule immediately implies that g is differentiable on I_1. We conclude that Figure 2 must belong to interval I_1 and that Figure 3 must belong to interval I_2.

2. (10 points) Compute the value of I_1 with a relative error which is less than $\tau = 10^{-9}$. Remember to explain why you can trust your error estimate!

   Hint Remember to determine and use the correct order p.

Solution We can rely on the error estimates as long as the fractions not only are close to 16, but are converging monotonically towards 16, i.e. the values corresponding to $k = 32, 64, 128, 256, 512$. In these cases a good error estimate is given by $(S_h - S_{2h})/15$. Therefore we look for a value of k for which

$$\left| \frac{\frac{S_h - S_{2h}}{15}}{\int_{I_1} g(x)dx} \right| < \tau = 10^{-9}$$

It is clear, that $0 < 0.989 < \int_{I_1} g(x)dx$ and the above inequality is certainly satisfied when

$$|S_h - S_{2h}| < 0.989 \times 15 \times 10^{-9} < 1.5 \times 10^{-8}$$

Scanning the fourth column of Figure 2 reveals that the value at $k = 64$ is sufficiently good, i.e.

$$\int_{I_1} g(x)dx \approx 9.89322425470486e - 01$$

with a relative error less that $\tau = 10^{-9}$.

Problem 3 Consider the function $h : [0, \infty) \to \mathbb{R}$ given by

$$h(x) = x^2 e^{-x} - \frac{1}{4}x.$$

1. (5 points) Explain why h has at least 3 distinct zeros on the interval $[0, \infty)$. You may rely on the following table of values of h.

x	h(x)
0.2200	-0.0162
0.4400	0.0147
0.6600	0.0601
0.8800	0.1012
1.1000	0.1278
1.3200	0.1355
1.5400	0.1234
1.7600	0.0929
1.9800	0.0463
2.2000	-0.0137

Solution We observe that the function h changes sign on the interval $[0.22, 0.44]$ and on the interval $[1.98, 2.20]$. It follows that h has at least one zero in each of these disjoint intervals. Moreover, since $h(0) = 0$, it is clear that there are at least three zeros in all.

2. (10 points) Newton's method has been used to compute a sequence of 10 approximations of a zero for h and these are the results

n	x(n)	h(x(n))
0	4.000000000000000e-01	7.251207365702311e-03
1	3.594915545716667e-01	3.365926732381980e-04
2	3.574094923949957e-01	1.050048316519892e-06
3	3.574029562463180e-01	1.043082287210950e-11
4	3.574029561813888e-01	-1.387778780781446e-17
5	3.574029561813890e-01	1.387778780781446e-17
6	3.574029561813888e-01	-1.387778780781446e-17
7	3.574029561813890e-01	1.387778780781446e-17
8	3.574029561813888e-01	-1.387778780781446e-17
9	3.574029561813890e-01	1.387778780781446e-17

Find a nonzero zero of h such that the relative error is less than $\tau = 10^{-6}$.

Solution There is little doubt that

$$\xi = 3.574029561814 \times 10^{-1}$$

is not a bad place to start looking for true root x^*. We want to be sure that ξ is such that

$$\frac{|x^* - \xi|}{|x^*|} < 10^{-6}.$$

There is not doubt that $\frac{1}{3} < x^*$ and so we merely have to ensure that

$$|x^* - \xi| < 3 \times 10^{-6} \tag{3}$$

To that end we compute $h(\xi \pm 3 \times 10^{-6})$. We have $h(\xi - 3 \times 10^{-6}) < 0$ and $h(\xi + 3 \times 10^{-6}) > 0$ so there is not doubt that the true root is between $\xi - 3 \times 10^{-6}$ and $\xi + 3 \times 10^{-6}$ which implies that the inequality (3) is satisfied. It follows that ξ is a good approximation of the true root x^* and that the relative error is less than $\tau = 10^{-6}$.

3. (10 points) When solving a general non-linear equation $h(x) = 0$ using Newton's method we iterate until $|h(x_n)| \leq \tau$ or we have completed `maxit` iterations. The parameters τ and `maxit` are specified by the user. Explain why it is usually pointless to use $\tau = 0$, despite the fact that we want to compute x such that $h(x) = 0$.

Solution If we choose $\tau = 0$, then we are really trying to obtain the exact solution. Suppose the computer returns a floating point number x for which $h(x)$ is evaluated as 0. Then there are at least two possibilities. The true root might actually be the floating point number x, but it is much more likely that the computation of $h(x)$ has under-flowed. Therefore it is likely that while x is probably a very good approximation, we have not accomplished our goal. Moreover, setting the tolerance too low is likely to result in the above behavior, where the residual is stagnating and no progress is made after a few iterations, wasting the users valuable time until `maxit` iterations have been completed.

Problem 4 Consider a ball which is being shot straight into the air at time $t = 0$. Let $y(t)$ denote the height of the ball above sea level at time t and let $v(t)$ denote the velocity. The ball is subject to gravity and air resistance and the motion of the ball is governed by the initial value problem

$$\begin{pmatrix} y'(t) \\ v'(t) \end{pmatrix} = \begin{pmatrix} f_1(y(t), v(t)) \\ f_2(y(t), v(t)) \end{pmatrix}, \quad \begin{pmatrix} y(0) \\ v(0) \end{pmatrix} = \begin{pmatrix} y_0 \\ v_0 \end{pmatrix},$$

where the two functions f_1 and f_2 are given by

$$f_1(y, v) = v, \quad f_2(y, v) = -g - \text{sign}(v)\frac{k}{m}v^2.$$

Here g is the constant of gravity, m is the mass of the ball and k is an aerodynamic constant.

1. (5 points) Let t^* denote the time when the ball reaches its maximum height above sea level. Show that $v(t^*) = 0$.

   Solution If y is *any* differentiable function and if t^* is the location of any local extrema, then $y'(t^*) = 0$. In our case, we have $y'(t) = v(t)$, so $v(t^*) = 0$.

2. (10 points) The trajectory of the ball has been approximated using Euler's explicit method and time step $h_1 = 0.1$. The results are given in Figure 4. Compute an approximation $t_1 \approx t^*$ using this data.

   Hint The fifth column is there to simplify your life. It contains the relevant values of f_2, i.e.

   $$f_2(n) = f_2(y_n, v_n).$$

   Solution Scanning column 4 we see the velocity change sign between t_7 and t_8. Since

   $$v_{n+1}(h) = v_n + hf_2(y_n, v_n)$$

   we see to determine h', such that

   $$0 = v_8(h') = v_7 + h'f_2(y_7, v_7)$$

   It follows that

   $$h' = -\frac{v_7}{f_2(y_7, v_7)} = \frac{0.5866}{9.8544} \approx 0.059526708881312$$

   Therefore, we pick

   $$t_1 = t_7 + h' \approx 0.7595$$

   as our approximation of t^*.

n	t(n)	y(n)	v(n)	f2(n)
0	0	0	10.0000	-19.8200
1	0.1000	1.0000	8.0180	-16.2488
2	0.2000	1.8018	6.3931	-13.9072
3	0.3000	2.4411	5.0024	-12.3224
4	0.4000	2.9414	3.7702	-11.2414
5	0.5000	3.3184	2.6460	-10.5201
6	0.6000	3.5830	1.5940	-10.0741
7	0.7000	3.7424	0.5866	-9.8544
8	0.8000	3.8010	-0.3988	-9.8041
9	0.9000	3.7611	-1.3793	-9.6298
10	1.0000	3.6232	-2.3422	-9.2714
11	1.1000	3.3890	-3.2694	-8.7511
12	1.2000	3.0621	-4.1445	-8.1023
13	1.3000	2.6476	-4.9547	-7.3651
14	1.4000	2.1521	-5.6912	-6.5810
15	1.5000	1.5830	-6.3493	-5.7886
16	1.6000	0.9481	-6.9282	-5.0200
17	1.7000	0.2553	-7.4302	-4.2992
18	1.8000	-0.4878	-7.8601	-3.6419
19	1.9000	-1.2738	-8.2243	-3.0561
20	2.0000	-2.0962	-8.5299	-2.5441

Figure 4: The approximate trajectory of the ball computed using Euler's explicit method and time step $h_1 = 0.1$.

3. (10 points) An even cruder approximation of the trajectory has also been computed using Euler's method and time step $h_2 = 2h_1 = 0.2$. The results are given in Figure 5. Use this data to estimate the accuracy of your approximation t_1 of t^*.

   Solution We observe v change sign between t_3 and t_4. Therefore we seek to determine h' such that

   $$0 = v_4(h') = v_3 + h' f_2(y_3, v_3)$$

   or equivalently

   $$h' = -\frac{v_3}{f_2(y_3, v_3)} = \frac{1.1558}{9.9536} \approx 0.116118791191127.$$

   Our second approximation of t^* becomes

   $$t_2 = t_3 + h' \approx 0.7116$$

n	t(n)	y(n)	v(n)	f2(n)
0	0	0	10.0000	-19.8200
1	0.2000	2.0000	6.0360	-13.4633
2	0.4000	3.2072	3.3433	-10.9378
3	0.6000	3.8759	1.1558	-9.9536
4	0.8000	4.1070	-0.8349	-9.7503
5	1.0000	3.9400	-2.7850	-9.0444
6	1.2000	3.3830	-4.5939	-7.7096
7	1.4000	2.4643	-6.1358	-6.0552
8	1.6000	1.2371	-7.3468	-4.4224
9	1.8000	-0.2323	-8.2313	-3.0445
10	2.0000	-1.8785	-8.8402	-2.0050

Figure 5: The approximate trajectory of the ball computed using Euler's explicit method and time step $h_2 = 0.2$.

In the past we have successfully applied an assumption of the type

$$t^* - t_1 = O(h^q)$$

only we do not have enough information to determine q. However, since the explicit Euler method is of order $p = 1$ in the time step h, it is extremely likely that $q = 1$ as in the case of the third project on the range of artillery guns. In this case the standard error estimate is $\frac{t_1 - t_2}{2^1 - 1} = t_1 - t_2 \approx 0.7595 - 0.7116 \approx 0.0479$. However, we do not have enough information to judge the quality of the error estimate so we are pushing the envelope here.