Problem 1 Let $f : \mathbb{R} \to \mathbb{R}$ denote the hyperbolic sine function given by

$$f(x) = \frac{e^x - e^{-x}}{2}. \tag{1}$$

In this problem you will consider the practical problem of computing f reliably on a binary computer using floating point arithmetic.

1. (5 points) Show that f is differentiable and strictly increasing.

   Solution f is form using the addition of two known differentiable functions and a single division. Hence f is differentiable. Moreover,

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

   so f is strictly increasing.

2. (5 points) Figure 1 shows the graph generated by the MATLAB commands

```
>> f=@(x)(exp(x)-exp(-x))/2;
>> x=single(linspace(-1,1,1025)*2^(-21));
>> plot(x,f(x));
```

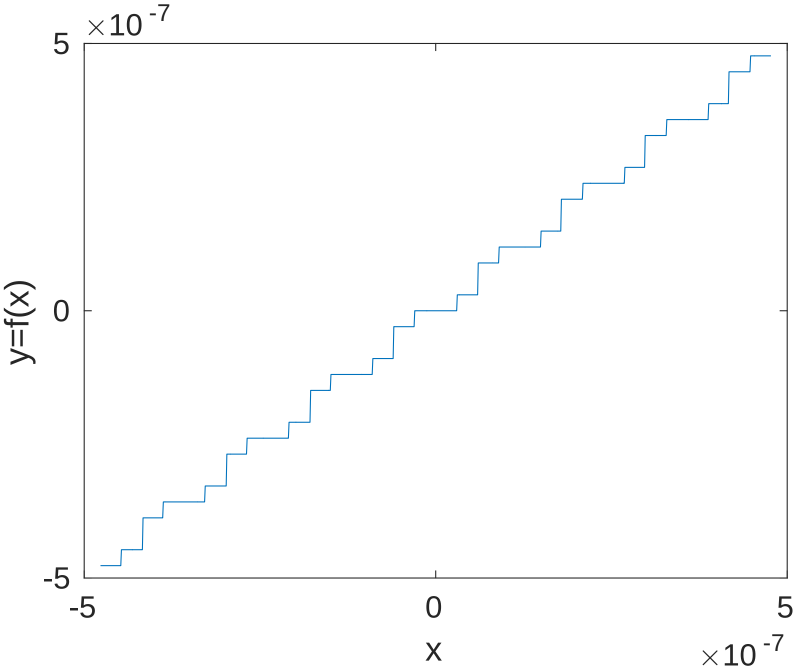

Figure 1: A naive attempt to construct the graph of f.

Why does Figure 1 immediately reveal that f has been computed in an unreliable manner?

Hint: There is more than one reason.

Solution The jumps suggests that this is not the graph of a continuous function. Moreover, there appears to be infinitely many points where f' is zero. Since f is differentiable, it is also continous. Moreover, we should have $f' > 0$ everywhere.

3. (5 points) Show that

$$f(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}. \tag{3}$$

Solution This is an immediate consequence of the known Taylor series

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \tag{4}$$

which implies that

$$e^x - e^{-x} = \sum_{n=0}^{\infty} \frac{x^n}{n!} - \sum_{n=0}^{\infty} (-1)^n \frac{x^n}{n!} = 2 \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} \tag{5}$$

as the even terms where $n = 2k$ cancel each other.

4. (10 points) Find a numerically reliable way to evaluate f accurately on the interval $(-2^{-21}, 2^{-21})$ using single precision arithmetic and explain why your approach has no risk of catastrophic cancellation.

Remark 1 Recall that the single precision round off error is $u = 2^{-24}$.

Solution It is a question of magnitude. We have

$$f(x) = x + \frac{1}{6}x^3 + O(x^5) \tag{6}$$

which suggests that we should use

$$f(x) \approx x \tag{7}$$

for x sufficiently small. Now the term $\frac{1}{6}x^3$ is only significant when

$$\frac{1}{6}|x|^3 > u|x| > 0 \tag{8}$$

This is certainly not the case for $x \in (-2^{-21}, 2^{-21})$ or even for the much larger interval $x \in (-2^{-12}, 2^{-12})$ where we have $x^2 \leq 2^{-24}$, so that

$$\frac{1}{6}|x|^3 < u|x| \tag{9}$$

The suggested approximation involves *no* arithmetic at all, hence there is zero risk of catastrophic cancellation.

Problem 2 Let $g : [0,1] \to \mathbb{R}$ be given by

$$g(x) = e^x \sin(x) \tag{10}$$

and consider the problem of computing the integral

$$T = \int_0^1 g(x)dx, \qquad \text{T : target}, \tag{11}$$

using the trapezoidal rule with uniform step size. Richardson's technique has been applied to the problem and Figure 2 contains all the results.

```
  k          Th                      Fh                     Eh
-----------------------------------------------------------------
  0    1.1436776435894214   0.00000000e+00   0.0000000000000000e+00
  1    0.9670583634015181   0.00000000e+00  -5.8873093395967767e-02
  2    0.9237047412420658   4.07392212e+00  -1.4451207386484088e-02
  3    0.9129205113631961   4.02009440e+00  -3.5947432929565779e-03
  4    0.9102279021357744   4.00512253e+00  -8.9753640914054988e-04
  5    0.9095549663092243   4.00128678e+00  -2.2431194218338243e-04
  6    0.9093867458976557   4.00032208e+00  -5.6073470522869741e-05
  7    0.9093446916415655   4.00008054e+00  -1.4018085363387556e-05
  8    0.9093341781304715   4.00002014e+00  -3.5045036980152489e-06
  9    0.9093315497560062   4.00000503e+00  -8.7612482176554118e-07
 10    0.9093308926625975   4.00000126e+00  -2.1903113622823156e-07
 11    0.9093307283892565   4.00000027e+00  -5.4757780310055182e-08
 12    0.9093306873209228   4.00000015e+00  -1.3689444577913434e-08
 13    0.9093306770538404   4.00000040e+00  -3.4223608021595928e-09
 14    0.9093306744870665   3.99999485e+00  -8.5559130151106422e-10
 15    0.9093306738453759   4.00001799e+00  -2.1389686318447806e-10
 16    0.9093306736849540   4.00001869e+00  -5.3473965995938975e-11
 17    0.9093306736448422   3.99936894e+00  -1.3370600922731532e-11
 18    0.9093306736348192   4.00198274e+00  -3.3409941480044836e-12
 19    0.9093306736323076   3.99058480e+00  -8.3721918286983055e-13
 20    0.9093306736316904   4.06961684e+00  -2.0572432646304151e-13
 21    0.9093306736315040   3.31089934e+00  -6.2135481944854590e-14
 22    0.9093306736314545   3.76457399e+00  -1.6505315632760660e-14
 23    0.9093306736314497   1.03720930e+01  -1.5913196686293911e-15
```

Figure 2: The results of integrating g numerically using the trapezoidal rule and many different values of the step size.

- The first column gives $k = \log_2(N)$ where N is the number of sub-intervals.
- Row k contains information pertaining to the step-size $h = 2^{-N}$.
- The second column contains the trapezoidal sums T_h.
- Richardson's fraction

  $$F_h = \frac{T_{2h} - T_{4h}}{T_h - T_{2h}} \tag{12}$$

  is the third column.
- Richardson's error estimate

  $$E_h = \frac{T_h - T_{2h}}{3} \tag{13}$$

  is in the fourth and final column.

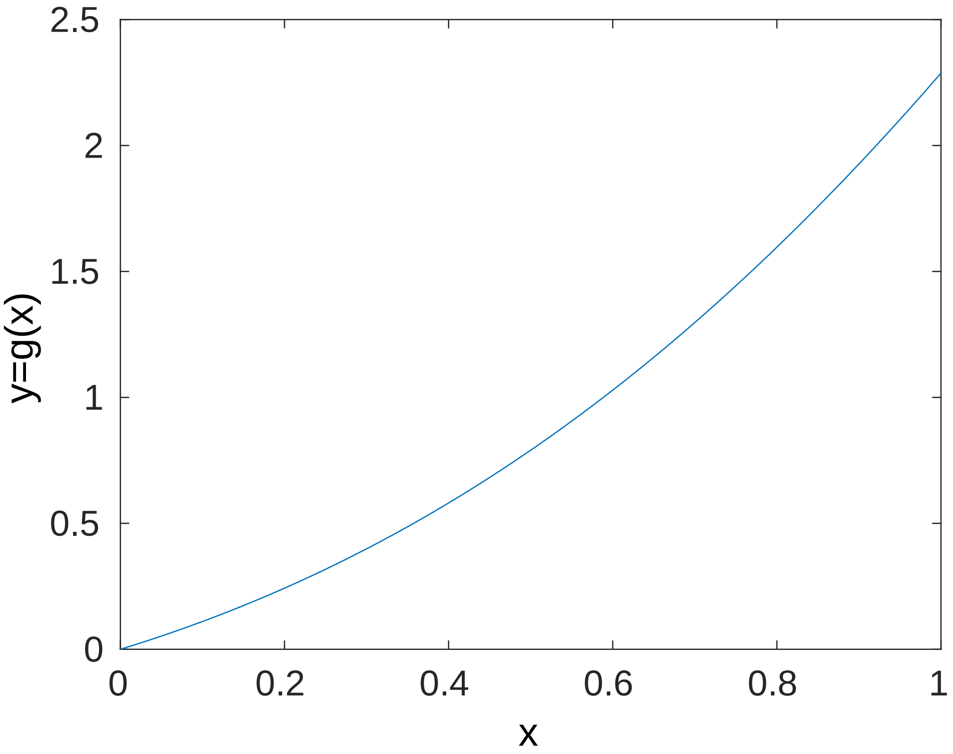

Figure 3: A reliable plot of the graph of g.

A reliable plot of the graph of g is given in Figure 3.

1. (5 points) While some of the error estimates are worthless they all have the same sign. Explain why the error $T - T_h$ will be negative if T_h is computed without any rounding errors at all.

   Hint: Take a good look at the graph of g and recall the geometric interpretation of the trapezoidal rule. Supplement with a short analysis of g and its derivatives as necessary.

   Solution The graph appears to be convex so the trapezoidal sum will overshot the target T and we get a negative error $E = T - A$. This is confirmed by a short calculation. We have

   $$g'(x) = e^x \sin(x) + e^x \cos(x) \tag{14}$$

   and

   $$g''(x) = e^x \sin(x) + e^x \cos(x) + e^x \cos(x) - e^x \sin(x) = 2e^x \cos(x) \tag{15}$$

   It is clear that $g''(x) > 0$ on the interval $[0, 1]$, so that function *is* convex.

2. (5 points) Find the first value of N where the computed value of Richardson's fraction deviates from the expected behavior.

   Solution This happens at $N = 12$. Until then the fractions have been converging monotonically towards 4 and the distance to 4 has been reduced by a factor of 4 from one row to the next. At $N = 12$, the fractions are still converging monotonically towards 4, but the speed is wrong.

3. (5 points) Explain why some people might believe that $N = 14$ is the correct answer to the previous question.

Solution At $N = 14$ the fractions have jumped to the other side of 4. This completely impossible for the true value of the fractions. This is something which is easy to detect, whereas tracking the speed with which the fractions approach 4 requires more insight and effort.

4. (10 points) Compute the value of T with a relative error which is less than $\tau = 10^{-7}$. Explain why you are confident that your estimate for the relative error is reliable.

   Solution We can use the values corresponding to $N = 11$. Here the error estimate is

   $$E_h = -5.4757780310055182 \times 10^{-8}. \tag{16}$$

   We trust the sign, magnitude an several digits of this error estimate, because the computed values of Richardson's fraction comply with the expected behavior at this point. In fact, based on our previous experience with cases where the exact error was available, we have reason to believe that the equality of the error estimate is maximal at this point. Moreover, it is clear that the target satisfies $T > 0.9$. It follows, that for $N = 11$ we have

   $$\frac{|T - T_h|}{|T|} < \frac{5.5 \times 10^{-8}}{0.9} \approx 6.1 \times 10^{-8} < 10^{-7}, \tag{17}$$

   so we should approximate the integral with

   $$T_h = 0.9093307283892565, \tag{18}$$

   although including all the digits would be overkill.

Problem 3 Consider the general problem of solving a nonlinear equation

$$h(x) = 0 \tag{19}$$

using an iterative method of your choice. Here $h : \mathbb{R} \to \mathbb{R}$ is infinitely often differentiable and has exactly one root ξ of multiplicity 1, i.e. $h(\xi) = 0$ and $h'(\xi) \neq 0$.

1. (5 points) Explain, how $h(x_k)$ may be extremely small, while x_k is nowhere near ξ.

   Solution The function h may converge asymptotically to zero as x tends to infinity, i.e.

   $$h(x) \sim e^{-x}, \quad x \to \infty. \tag{20}$$

   In this case, Newton's method yields

   $$x_{n+1} = x_n - \frac{h(x_n)}{h'(x_n)} \sim x_n + 1. \tag{21}$$

   which implies

   $$x_n \to \infty, \quad n \to \infty \tag{22}$$

   and

   $$h(x_n) \to 0 \tag{23}$$

   despite the fact that we are nowhere near a root.

2. (5 points) Explain, how the difference between successive approximation $|x_{k+1} - x_k|$ can be extremely small, while x_k is nowhere near ξ.

   Solution We may have

   $$h(x) \sim e^{-\lambda x}, \quad x \to \infty. \tag{24}$$

   In this case, Newton's method yields

   $$x_{n+1} = x_n - \frac{h(x_n)}{h'(x_n)} \sim x_n + \frac{1}{\lambda}. \tag{25}$$

   which implies

   $$x_{n+1} - x_n \sim \frac{1}{\lambda} \tag{26}$$

   can be arbitrarily small, depending on the value of λ, despite the fact that we can not converge.

3. (7 points) Explain the value of computing brackets (a_k, b_k) such that $h(a_k)h(b_k) < 0$ and $|b_k - a_k| \to 0$ for $k \to \infty$.

   Solution If we have interval (a_k, b_k) such that $h(a_k)h(b_k) < 0$ then we how that h changes sign on the interval. By the continuity of h this implies that h has a zero in this interval. We can use the midpoint c_k as an existimate of the root and the error will be at most $\frac{1}{2}|b_k - a_k|$. As long as we can trust the computed sign at the endpoints, we can obtain better and better approximations of the root and reliable error estimates.

4. (8 points) Suppose that all calculations are done in floating point arithmetic. Explain, why the happy occurrence that $h(x_k)$ is evaluated as 0, should not be taken as evidence that x_k is exactly equal to ξ.

   Solution It is almost certain that rounding errors are made when computing $h(x_k)$. Hence, the computed value of $h(x_k)$ can equal 0 although $x_k \neq \xi$.

Problem 4 A trajectory for an artillery shell has been carefully computed using an unknown method and different values of the time step h. Richardson's techniques have been applied and all relevant information about the trajectory is given in Figures 4, 5, 6, and 7. As usual, the shell starts at $(0, 0)$ and is fired in the direction of the positive x axis. As usual, gravity acts parallel to the y axis and pulls the shell downward.

1. (5 points) What evidence do you find to support the conjecture that the shell is moving through an atmosphere?

   Solution In vacuum, the x velocity would be constant. By examing Figure 6 we see that the shell x component is being significantly reduced.

2. (5 points) What evidence do you find to support the conjecture that the trajectory has been computed using a method for which the global error is $O(h^3)$ where h is the time step size?

   Solution We scanning all rows of all tables we see many instances where Richardson's fraction is moving monotonically towards $8 = 2^3$ as the step size is reduced. This supports the conjecturat the order is $p = 3$. There are cases where the factions are executing an illegal jump to the other side of 8, but they are still close to 8. Such jumps are the result of catastropic cancellation in the computation of the nominator and denominator of the factions. This is unavoidable and a prominant feature of methods of relatively high order. Our previous experience suggests that artillery trajectories are smooth, i.e. infinitely often differentiable, so we have no reason to believe that the order would not be an integer.

3. (5 points) What evidence do you find to support the conjecture that the shell is aimed at a target located at $(7000, 0)$ and that a real shell would destroy an (unarmored) target located there?

   Solution The shell appears to touching the ground immediately after the 10th time step, see the last row Figure 5 where the shell is around 1 centimer above the ground. Here the x coordinate rounds to 7000 m with 4 significant figures. We have no reason to doubt these figures as Richardson's fraction are close to 8. The kill radius of our standard shell is 15 m against an unarmoured target, so the is no doubt that the shell would destroy a target at 7000 m.

4. (10 points) Why can we say with certainty that the shell reaches the highest point of its trajectory roughly 6 seconds after leaving the muzzle of the gun?

   Solution This is a question of solving $y'(t) = 0$. We detect a sign change for the compute approximation of y' between the 4th and the 5th timestep. We trust the magnitude of the the error estimate for these two values because Richardson's fraction are close to 8. This mean that the true value of y' changes sign in this interval as well. By the continuity of y', there is a zero in this interval. It is clear from the values of y' that the zero is (probably) closer to the right hand endpoint of the interval, i.e. close to 5.8 seconds. This is roughly 6 seconds after the shell has left the muzzle of the gun.

   Remark 2 Notice that this question is valued at 10 points. To get full credit you will have to argue why every single number you reference is trustworthy to the extent which you require for your arguments to be valid.

```
   t          x_h(t)          F_4h(t)         F_2h(t)         F_h(t)          E_h(t)
------------------------------------------------------------------------------------------
 0.000000e+00  0.00000000e+00             NaN             NaN             NaN  0.00000000e+00
 1.162263e+00  8.76094533e+02  8.34392699e+00  8.16877044e+00  8.08378819e+00 -8.24670813e-06
 2.324526e+00  1.70144667e+03  8.32016787e+00  8.15807091e+00  8.07870182e+00 -1.42605950e-05
 3.486790e+00  2.48165346e+03  8.29797811e+00  8.14808252e+00  8.07395649e+00 -1.86593398e-05
 4.649053e+00  3.22142548e+03  8.27706548e+00  8.13867842e+00  8.06949250e+00 -2.18736727e-05
 5.811316e+00  3.92476296e+03  8.25720131e+00  8.12975853e+00  8.06526258e+00 -2.42092704e-05
 6.973579e+00  4.59509018e+03  8.23820452e+00  8.12124299e+00  8.06122890e+00 -2.58864566e-05
 8.135843e+00  5.23535942e+03  8.21993010e+00  8.11306746e+00  8.05736095e+00 -2.70662404e-05
 9.298106e+00  5.84813244e+03  8.20226084e+00  8.10517957e+00  8.05363387e+00 -2.78677396e-05
 1.046037e+01  6.43564506e+03  8.18510112e+00  8.09753644e+00  8.05002698e+00 -2.83800520e-05
 1.162263e+01  6.99985882e+03  8.16837234e+00  8.09010256e+00  8.04652339e+00 -2.86704747e-05
```

Figure 4: Data related to the x-coordinate of the shell.

```
   t          y_h(t)          F_4h(t)         F_h(t)          F_h(t)          E_h(t)
------------------------------------------------------------------------------------------
 0.000000e+00  0.00000000e+00             NaN             NaN             NaN  0.00000000e+00
 1.162263e+00  6.45662404e+01  8.11995685e+00  8.06265232e+00  8.03218926e+00 -1.15052119e-06
 2.324526e+00  1.12507680e+02  8.09089311e+00  8.04971071e+00  8.02606818e+00 -2.03242877e-06
 3.486790e+00  1.44920697e+02  8.06622164e+00  8.03877559e+00  8.02090897e+00 -2.72984544e-06
 4.649053e+00  1.62729111e+02  8.04544653e+00  8.02961870e+00  8.01660138e+00 -3.29879992e-06
 5.811316e+00  1.66718926e+02  8.02807993e+00  8.02201467e+00  8.01303665e+00 -3.77712767e-06
 6.973579e+00  1.57564891e+02  8.01365113e+00  8.01574608e+00  8.01010995e+00 -4.19067116e-06
 8.135843e+00  1.35851114e+02  8.00171668e+00  8.01060850e+00  8.00772285e+00 -4.55724428e-06
 9.298106e+00  1.02087271e+02  7.99186872e+00  8.00641434e+00  8.00578521e+00 -4.88921389e-06
 1.046037e+01  5.67215101e+01  7.98374043e+00  8.00299537e+00  8.00421635e+00 -5.19520874e-06
 1.162263e+01  1.50839035e-01  7.97700855e+00  8.00020370e+00  8.00294539e+00 -5.48126714e-06
```

Figure 5: Data related to the y coordinate of the shell.

```
    t          x'_h(t)        F_4h(t)        F_2h(t)        F_h(t)         E_h(t)
-----------------------------------------------------------------------------------
 0.000000e+00  7.77446486e+02            NaN            NaN            NaN  0.00000000e+00
 1.162263e+00  7.31078833e+02  8.77590525e+00  8.36025272e+00  8.17351231e+00  4.19329581e-07
 2.324526e+00  6.89970363e+02  8.75286071e+00  8.35019552e+00  8.16882549e+00  6.82624692e-07
 3.486790e+00  6.53267311e+02  8.73309525e+00  8.34152540e+00  8.16477453e+00  8.43479565e-07
 4.649053e+00  6.20289724e+02  8.71604866e+00  8.33401423e+00  8.16125665e+00  9.36351528e-07
 5.811316e+00  5.90489303e+02  8.70128355e+00  8.32748228e+00  8.15819087e+00  9.83798584e-07
 6.973579e+00  5.63418941e+02  8.68845447e+00  8.32178668e+00  8.15551248e+00  1.00081744e-06
 8.135843e+00  5.38710360e+02  8.67728579e+00  8.31681257e+00  8.15316947e+00  9.97507479e-07
 9.298106e+00  5.16057426e+02  8.66755562e+00  8.31246696e+00  8.15111941e+00  9.80742470e-07
 1.046037e+01  4.95203545e+02  8.65908386e+00  8.30867401e+00  8.14932807e+00  9.55239898e-07
 1.162263e+01  4.75932004e+02  8.65172327e+00  8.30537140e+00  8.14776633e+00  9.24257214e-07
```

Figure 6: Data related to the x component of the velocity of the shell.

```
    t          y'_h(t)        F_4h(t)        F_2h(t)        F_h(t)         E_h(t)
-----------------------------------------------------------------------------------
 0.000000e+00  6.30631576e+01            NaN            NaN            NaN  0.00000000e+00
 1.162263e+00  4.82288764e+01  7.33697610e+00  7.77759697e+00  7.91241324e+00 -2.31963751e-08
 2.324526e+00  3.44243859e+01  7.63232155e+00  7.88800270e+00  7.95992838e+00 -5.02648995e-08
 3.486790e+00  2.14832854e+01  7.79019506e+00  7.94811684e+00  7.98597643e+00 -7.72244039e-08
 4.649053e+00  9.27342264e+00  7.88408274e+00  7.98411036e+00  8.00159476e+00 -1.02075057e-07
 5.811316e+00 -2.31136169e+00  7.94329575e+00  8.00679145e+00  8.01141103e+00 -1.23911525e-07
 6.973579e+00 -1.33571946e+01  7.98174133e+00  8.02140215e+00  8.01769147e+00 -1.42429599e-07
 8.135843e+00 -2.39345479e+01  8.00681510e+00  8.03077239e+00  8.02166766e+00 -1.57648342e-07
 9.298106e+00 -3.41015013e+01  8.02277929e+00  8.03655084e+00  8.02406057e+00 -1.69752605e-07
 1.046037e+01 -4.39062095e+01  8.03223970e+00  8.03975357e+00  8.02531594e+00 -1.79003860e-07
 1.162263e+01 -5.33887916e+01  8.03686120e+00  8.04103368e+00  8.02572120e+00 -1.85690237e-07
```

Figure 7: Data related to the y component of the velocity of the shell.