Problem 1 The trajectory of an artillery shell has be carefully computed and some of the results are given in Table 1.

t	x(t)	y(t)	x'(t)	y'(t)
s	m	m	m/s	m/s
0	0	0	5.5154e+02	5.5154e+02
5	2.4423e+03	2.3287e+03	4.3549e+02	3.9151e+02
10	4.4263e+03	3.9967e+03	3.6321e+02	2.8147e+02
15	6.1129e+03	5.1866e+03	3.1433e+02	1.9775e+02
20	7.5925e+03	5.9994e+03	2.7928e+02	1.2931e+02
25	8.9200e+03	6.4952e+03	2.5288e+02	7.0279e+01
30	1.0130e+04	6.7123e+03	2.3199e+02	1.7390e+01
35	1.1246e+04	6.6763e+03	2.1452e+02	-3.1169e+01
40	1.2279e+04	6.4064e+03	1.9904e+02	-7.6230e+01
45	1.3238e+04	5.9195e+03	1.8458e+02	-1.1798e+02
50	1.4125e+04	5.2324e+03	1.7062e+02	-1.5625e+02
55	1.4944e+04	4.3632e+03	1.5692e+02	-1.9079e+02
60	1.5695e+04	3.3310e+03	1.4350e+02	-2.2142e+02
65	1.6380e+04	2.1555e+03	1.3046e+02	-2.4812e+02
70	1.7001e+04	8.5618e+02	1.1795e+02	-2.7100e+02
T	1.7353e+04	0	1.1055e+02	-2.8335e+02

Figure 1: The trajectory of a shell fired at time $t = 0$ until the time of impact $t = T = 70.3088$ s.

1. (5 points) Which property about the trajectory must either be proved or assumed before it follows with certainty that the shell reaches its heighest point some time between 30 and 35 seconds after being fired?

   Solution If we know that y' is continuous, then it follows with certainty that $y'(\tau) = 0$ for at least one $\tau \in (30, 35)$, simply because $y'(30)$ and $y'(35)$ have different signs. It is the nature of shell trajectories that there is one and only one point where y is maximal. In short, if we know that y is differentiable and y' is continuous, then information in the table allows us to conclude that the shell reaches its highest point for some $\tau \in (30, 35)$.

2. (5 points) Compute the total work done by the friction between the shell and the surrounding atmosphere from time $t = 0$ until the time of impact $t = T$ *relative* to the initial kinetic energy of the shell.

   Solution The work done by friction equals the loss of total mechanical energy. The potential energy, i.e. $\phi = \phi(y) = mgy$ is zero, when the shell is fired and when the shell impacts. Hence, we need only compute the kinetic energy when the shell is launched and when the shell impacts. It

is irrelevant that we do not know the mass of the shell, because we need to estimate the work done by friction *relative* to the initial energy. We find the result to be

$$\rho = \frac{\frac{1}{2}mv_0^2 - \frac{1}{2}mv_T^2}{\frac{1}{2}mv_0^2} = \frac{v_0^2 - v_T^2}{v_T^2} \approx 0.8479 \tag{1}$$

indicating that the shell has lost about 85% of the energy imparted to it by the force of the exploding propellant.

3. (15 points) The shell is at height $y = 6000$ meters twice during its flight. Pick one of the two points and compute the corresponding value of the x coordinate as accurately as you can. Estimate the relative error on your result.

Solution It follows from the continuity of the trajectory that the shell is at height $y = 6000$ meters at some time $\tau \in (40, 45)$. We have $y(40) = 6406.4$ and $y(45) = 5919.5$. The simplest thing to do is to interpolate linearly between these two points. We seek ν, such that

$$6000 = \nu y(40) + (1 - \nu)y(45) = y(45) + \nu(y(40) - y(45)). \tag{2}$$

This equation has the unique solution

$$\nu = \frac{6000 - y(45)}{y(40) - y(45)} = 0.1653 \tag{3}$$

which we then use to estimate the relevant x coordinate

$$x \approx \nu x(40) + (1 - \nu)x(45) = 13079 \tag{4}$$

Now as for the relative error we can say with certainty that $x \in (x(40), x(50)$. It follows that our error is certainly bounded by $x(45) - x(40) \approx 951$ and the relative error is bounded by $(x(45) - x(40))/x(40) \approx 0.0781$. However, in view of the fact that $y(45)$ is much closer to 6000, than $y(40)$, our estimate is almost certainly much better than this analysis would indicate.

Problem 2 An unknown function $f : [0, 1] \to \mathbb{R}$ has been integrated numerically using the trapezoid rule T_h where $h = 1/N$ and $N = 2^k$. The computed results are given as Table 2. The table also contains Richardson's fractions $F_h = \frac{T_{2h}-T_{4h}}{T_h-T_{2h}}$ and his error estimates $E_h = \frac{T_h-T_{2h}}{3}$.

k	Th	fraction Fh	Eh
0	1.1436776435894214	0.00000000e+00	0.0000e+00
1	0.7757889068338907	0.00000000e+00	-1.2263e-01
2	0.6901947335054285	4.29805818e+00	-2.8531e-02
3	0.6695218208542304	4.14040221e+00	-6.8910e-03
4	0.6644028353205115	4.03847843e+00	-1.7063e-03
5	0.6631262165469027	4.00979967e+00	-4.2554e-04
6	0.6628072580636909	4.00246064e+00	-1.0632e-04
7	0.6627275307173217	4.00061582e+00	-2.6576e-05
8	0.6627075996480573	4.00015400e+00	-6.6437e-06
9	0.6627026169287019	4.00003850e+00	-1.6609e-06
10	0.6627013712518608	4.00000963e+00	-4.1523e-07
11	0.6627010598328377	4.00000240e+00	-1.0381e-07
12	0.6627009819780926	4.00000055e+00	-2.5952e-08
13	0.6627009625144066	4.00000006e+00	-6.4879e-09
14	0.6627009576484860	4.00000078e+00	-1.6220e-09
15	0.6627009564320073	4.00000475e+00	-4.0549e-10
16	0.6627009561278847	3.99996130e+00	-1.0137e-10
17	0.6627009560518562	4.00011244e+00	-2.5343e-11
18	0.6627009560328556	4.00137897e+00	-6.3335e-12
19	0.6627009560280878	3.98514379e+00	-1.5893e-12
20	0.6627009560268888	3.97675711e+00	-3.9964e-13
21	0.6627009560266003	4.15505964e+00	-9.6182e-14
22	0.6627009560265480	5.51804671e+00	-1.7431e-14
23	0.6627009560264855	8.36589698e-01	-2.0835e-14

Figure 2: The results obtained by integrating an unknown function numerically using $N = 2^k$ subintervals along with some auxiliary numbers explained in the main text.

1. (5 points) The function f is costly to compute and it requires a full CPU second to evaluate a f at a single point x. Assuming that the values are recycled how many CPU days would it require to compute the above table from scratch?

   Solution On the finest grid there are $2^{23} + 1$ function points were the function f has to be evaluated. Since the grids are formed by repeated subdivision, the function values can be recycled. The price of doing the function evaluations is the $2^{23} + 1 = 8388609$ CPU seconds. This is about

97 CPU days, a nontrivial amount of computing time.

2. (10 points) Consider the value of Richardson's fraction corresponding to $k = 14$, i.e. $F_h = 4.00000078$. What conclusions can be drawn about this value which would be impossible to reach if you did not have the rest of the table available?

   Solution If we ignore the rest of the table, then there is very little which can be said. Given that we are dealing with the trapezoidal rule we would suspect that the function is differentiable and that we are very close to the point were rounding errors become significant. If we use the entire table, then we see Richardson's fractions converge monotonically down to 4 from $k = 2$ to $k = 13$. This is exactly the kind of behaviour which we would expect from applying the trapezoidal rule to a smooth function. Moreover, we can say with certainty that rounding errors are starting to become significant, because the value of Richardsons fraction at $k = 14$ has moved away from 4. In view of our past experience we can now say that we are starting to loose significant figures in the error estimate and that there is no point in decreasing h further.

3. (10 points) Estimate the integral with a a relative error which is less than $\tau = 10^{-8}$ and explain why your error estimate is realiable.

   Solution At $k = 13$ the absolute error estimate is $E_h \approx -6.4879 \times 10^{-9}$. This error estimate is reliable, because we are inside the range where the computed values of Richardson's fraction are converging monotonically towards 4. In fact we are exactly at the turning point, so we expect that not only is the sign and the order of magnitude correct, but several significant figures as well. It is clear from the second column that the integral is strictly larger than 6.6×10^{-1}. It follows that the value of T_h for $k = 13$ approximates the integral with a relative error which is less than

$$\frac{-6.4879 \times 10^{-9}}{6.6 \times 10^{-1}} < 10^{-8} = \tau \tag{5}$$

Problem 3 Consider the problem of computing the function

$$f(x) = \sinh(x) = \frac{e^x - e^{-x}}{2} \tag{6}$$

from scratch using floating point arithmetic.

1. (5 point) Compute the condition number $\kappa_f(x)$ of f for all $x \neq 0$.

   Solution By definition, we have

   $$\kappa_f(x) = \left| \frac{xf'(x)}{f(x)} \right| = \left| \frac{x\cosh(x)}{\sinh(x)} \right| = \frac{x\cosh(x)}{\sinh(x)} \tag{7}$$

   for all $x \neq 0$. The absolute value signs are not necessary, because $\cosh(x) > 0$ and because x and $\sinh(x)$ always have the same sign.

2. (10 point) Explain why it is theoretically possible to compute $f(x)$ with a relative error which is less than $2u$ where u is the unit round off error for all x in the interval $[-1, 1]$.

   Solution In reality, it is a question of showing that the condition number is bounded by 2. By l'Hospital's theorem we have $\kappa_f(x) \to 1$ as $x \to 0$, which permits us to extend $\kappa_f(x)$ continuously to the point $x = 0$. Moreover, since

   $$\begin{aligned} \frac{d}{dx}\kappa_f(x) &= \frac{(\cosh(x) + x\sinh(x))\sinh(x) - x\cosh(x)^2}{\sinh(x)^2} \\ &= \frac{\cosh(x)\sinh(x) + x(\sinh(x)^2 - \cosh(x)^2)}{\sinh(x)^2} = \frac{\cosh(x)\sinh(x) - x}{\sinh(x)^2} \end{aligned}$$

   is an *odd* function, we deduce that the maximum of $\kappa_f(x)$ on the interval $[-1, 1]$ is assumed either at $x = 0$ or at $x = \pm 1$. Now, since $\kappa_f(\pm 1) \approx 1.3$ there is plenty of wriggle room.

3. (10 point) Explain why the definition of f is unsuitable for direct numerical computation for x sufficiently close to 0 and explain in detail how to ensure that the relative error is less than a given tolerance τ.

Solution The naive expression is unsuitable for direct numerical computation, because it will cancel catastropically for $x \approx 0$. In order to derive a good algorithm we notice that $f(x) = \sinh(x)$ is an *odd* function. It follows that it is enought to compute x for positive values of x. Moreover, since $\exp(-x) = \exp(x)$ it is enough to compute $\exp(x)$. The expression

$$g(x) = \frac{e^x - \frac{1}{e^x}}{2} \tag{8}$$

will of course still cancel catastropically for x sufficiently small, but there will not be any problems for $x > \frac{\log(2)}{2}$, because

$$e^x > 2e^{-x} \Leftrightarrow e^{2x} > 2 \Leftrightarrow 2x > \log(2) \Leftrightarrow \frac{\log(2)}{2} \tag{9}$$

and we know that any subtraction $d = a - b$ is safe when $a > 2b$. It remains to handle the interval $[0, \frac{log(2)}{2}]$. If we approximate $f(x)$ with the Taylor polynomial of order $2m + 1$, then

$$f(x) - p_{2m+1}(x) = \frac{1}{(2m+2)!} f^{(2m+2)}(\xi) x^{2m+2} \tag{10}$$

where $\xi \in (0, x)$ and since $f^{(2m+2)}(x) = f(x)$ we have

$$\begin{aligned} 0 \leq \frac{f(x) - p_{2m+1}(x)}{f(x)} = \frac{1}{(2m+2)!} \frac{f(\xi)}{f(x)} x^{2m+2} &\leq \frac{1}{(2m+2)!} x^{2m+2} \\ &\leq \frac{1}{(2m+2)!} \left(\frac{\log(2)}{2} \right)^{2m+2} \to 0, \quad m \to \infty. \end{aligned}$$

The second to last inequality hinges on the fact that f is monotone increasing for $x \geq 0$. It is now clear that the relative error can be made as small as desired simply by increasing m. In particular if $m = 6$, then the relative error estimate for p_{2m+1} is roughly 4.1374×10^{-18} for the entire interval $[0, \log(2)/2]$, well below the unit round off error in double precision. Thus we approximate f with p_{13} on the interval $[0, \log(2)/2]$. We use $g(x)$ for $x > \log(2)/2$ and we exploit the fact that $f(x) = -f(-x)$ for $x < 0$.

Problem 4 Consider the problem of computing the reciprocal square root of a postive real number α, i.e. $\alpha \to 1/\sqrt{\alpha}$.

1. (5 points) Show that $x = 1/\sqrt{\alpha}$ if and only if x is the solution of the equation
$$f(x) = 0, \tag{11}$$
where $f(x) = \frac{1}{x^2} - \alpha$.

   Solution Since $x > 0$ it follow that
$$x = \frac{1}{\sqrt{\alpha}} \Leftrightarrow \frac{1}{x} = \sqrt{\alpha} \Leftrightarrow \frac{1}{x^2} - \alpha = 0 \Leftrightarrow f(x) = 0. \tag{12}$$

2. (10 points) Of all the basic arithmetic operations divisions has always been the slowest in terms of CPU cycles. Show that Newton's method can be applied to computing the reciprocal square root of α without doing any divisions.

   Solution We experimentally apply Newton's method to the nonlinear equation $f(x) = 0$. We have
$$f'(x) = -2x^{-3} \tag{13}$$
and it follows that Newton's method can be written as
$$\begin{aligned} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} &= x_n - \frac{x_n^{-2} - \alpha}{-2x_n^{-3}} \\ &= x_n + \frac{1}{2}(x_n - \alpha x_n^3) = x_n + \frac{1}{2}x_n(1 - \alpha x_n^2). \end{aligned}$$
   This last expression does not require any floating point divisions.

3. (10 points) The proper use of Newton's method requires the construction of a good initial guess. This can be very difficult if α is an arbitrary positive number in the representable range. However, any positive floating point number α admits a binary representation of the form
$$\alpha = (1.f_1 f_2 f_3 .... f_k)_2 \times 2^m, \quad f_k \in \{0, 1\} \tag{14}$$
where $k > 0$ and m are integers. Explain why it suffices to construct a good initial guess for $\alpha \in [1, 4]$ and why $x_0(\alpha) = \frac{3}{4}$ is not a bad initial guess for $1/\sqrt{\alpha}$ for all $\alpha \in [1, 4]$.

   Solution We have
$$\begin{aligned} \sqrt{\alpha} &= \sqrt{(1.f_1 f_2 f_3 .... f_k)_2 \times 2^m} \\ &= \begin{cases} \sqrt{(1.f_1 f_2 f_3 .... f_k)_2 \times 2} \times 2^k & m = 2k + 1 \\ \sqrt{(1.f_1 f_2 f_3 .... f_k)_2} \times 2^k & m = 2k \end{cases} \end{aligned}$$

This shows that we only need to worry about the square root of α for $\alpha \in [1, 4]$. In this interval we have $\sqrt{\alpha} \in [1, 2]$ and $1/\sqrt{\alpha} \in [1/2, 1]$. It follows that the midpoint of this last interval is not a bad initial guess. Certainly, we could do better, but that is not the question here.