Problem 1 Consider the function $f : (0, \infty) \to 0$ given by

$$f(x) = \sqrt{9 + x^2} - 3. \tag{1}$$

1. (6 points) Show that f is a strictly positive and strictly increasing function of x.

   Solution

   - (Strictly increasing, 3pt.) The function is built from functions which are all differentiable using elementary arithmetic operations and function compostion. Hence, it is differentiable, and

     $$f'(x) = (9 + x^2)^{-\frac{1}{2}} x > 0, \quad x \in (0, \infty). \tag{2}$$

     It follows that f is strictly increasing.

   - (Strictly positive, 3pt.) Let $x \neq 0$. Then

     $$f(x) = \sqrt{9 + x^2} - 3 > \sqrt{9} - 3 = 0 \tag{3}$$

     because f is strictly increasing

2. (6 points) The following MATLAB commands

   ```
   >>f=@(x)sqrt(9+x.^2)-3;
   >>x=single(linspace(0,1,1025)*2^-8);
   >>plot(x,f(x));
   ```

   have produced the figure given in Figure 1. What evidence do you find to support the statement that this is not a reliable way to compute f.

   Solution There are two observations which can be made

   - (3pt) The function is *not* strictly positive, but appears to be identical to zero for all x in the interval $(0, \rho)$, for some $\rho > 0$. The exact value of ρ can not be determined from the graph, but it is clear that $\rho > \frac{1}{2}$ so the interval is substantial.
   - (3pt) The function is *not* strictly increasing, but appears to be piecewise constant.

   These observed properties of the naive implementation are in direct violation of the known properties of the function f. Therefore, it is doubtful in the extreme that the naive implementation is reliable.

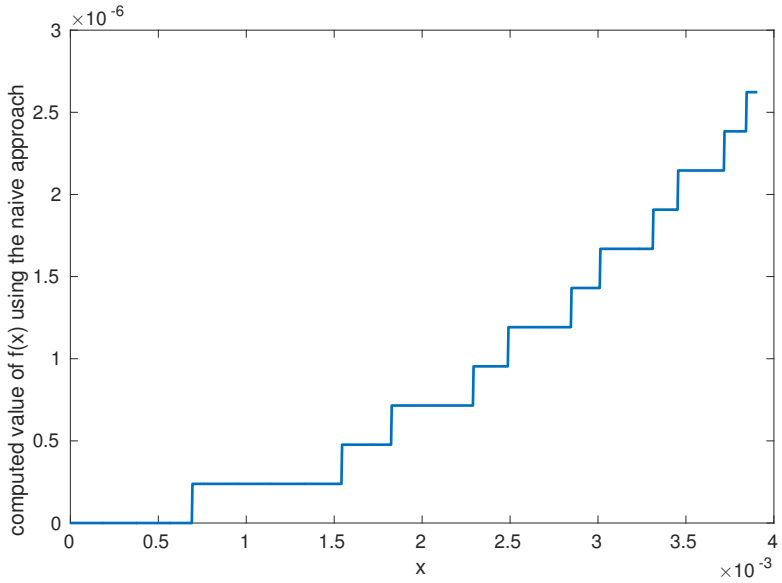

Figure 1: A plot of the computed values of f using the naive implementation based directly on the definition of f.

3. (5 points) Explain why the *computed* value of f equals zero for all $x \leq \frac{1}{2} \times 10^{-3}$ while the true value of f is strictly positive for all $x > 0$.

   Solution Let $a = (1.f_1 f_2 \dots f_{23})_2 \cdot 2^m$ denote a positive single precision number. The next single precision number is

   $$a_+ = \left[(1.f_1 f_2 \dots f_{23})_2 + 2^{-23}\right] \cdot 2^m \tag{4}$$

   For any *real* number $\xi \in (a, a + 2^{m-24})$, we have

   $$\text{fl}(\xi) = a \tag{5}$$

   because a is the *closest* floating point number. In particular, when

   $$a = 9 = 1.125 \cdot 2^3 = (1.001)_2 \cdot 2^3 \tag{6}$$

   it follows that

   $$\text{fl}(9 + x^2) = 9 \tag{7}$$

   when $x^2 < 2^{-21}$ or equivalently $|x| < 2^{-\frac{21}{2}} \approx 6.9053 \cdot 10^{-4}$. In particular,

   $$\text{fl}(9 + x^2) = 9, \quad 0 < x < \frac{1}{2} \cdot 10^{-3}. \tag{8}$$

4. (8 points) Find a reliable way to evaluate f and explain why it can *never* cancel catastrophically.

Solution

- (The equivalent expression, 4pts) We have

$$f(x) = \sqrt{9+x^2} - 3 = (\sqrt{9+x^2} - 3)\frac{\sqrt{9+x^2}+3}{\sqrt{9+x^2}+3} = \frac{x^2}{\sqrt{9+x^2}+3} \quad (9)$$

- (The absence of subtractive cancellation, 4pts) The new expression

$$f(x) = \frac{x^2}{\sqrt{9+x^2}+3} \quad (10)$$

is *mathematically* equivalent to the definition, but it does not contain *any* subtractions, so catastropic cancellation is completely impossible.

elevation	wind speed (meters/second)				
(degrees)	-2	-1	0	1	2
0	0	0	0	0	0
10	11424	11438	11453	11467	11481
20	15788	15817	15887	15876	15906
30	17586	17628	17670	17712	17754
40	17708	17759	17810	17861	17913
50	16458	16515	17773	16631	16688
60	13965	14026	14087	14148	14209
70	10293	10354	10415	10476	10538
80	5518	5577	5635	5694	5753
90	-113	-56	-2	56	113

Figure 2: A partial artillery table for a new piece of artillery.

Problem 2 Your brigade has received a new piece of artillery along with the very partial firing table show in Figure 2. It gives the range as a function of the elevation of the gun and the wind speed parallel to the ground. Negative values of the wind indicate head wind (Swedish: "motvind"), positive wind speeds indicate tail wind (Swedish: "medvind").

1. (6 points) Exactly three entries in the column corresponding to $w = 0$ have been corrupted. Find them and explain why they are wrong.

   Solution There are three entries to find and three explanations to make, i.e. six things to do. Let θ denote the elevation of the gun and let w denote the wind

   - The value corresponding to $\theta = 20°$ is wrong, because the range $r = 15887$ is strictly larger than the range corresponding to $\theta = 20°$ and $w = 1$ (tail wind). Obviously, the range should increase when there is tail wind.
   - The value corresponding to $\theta = 50°$ is wrong, because the range $r = 17773$ is strictly larger than the range corresponding to $\theta = 50°$ and $w = 1$ (tail wind). Obviously, the range should increase when there is tail wind.
   - The value correspoinding to $\theta = 90°$ is wrong, because the range $r = -2$ is negative. In our model, which does not incorporate the rotation of the Earth, the shell should fall back into the barrel and not land behind the gun, unless there is head wind.

2. (4 points) Assuming the errors are confined to the column $w = 0$, find a way to approximate the correct value of the corrupted data accurately.

   Solution For each value of θ, we can interpolate linearly using the range corresponding to $w = -1$ and $w = 1$. The actual computation is trivial, because $w = 0$ is in the middle of the interval $(-1, 1)$, so it suffice to compute an average of two known values. For instance, in order to repair the flawed value of r corresponding to $\theta = 20°$, we simply compute the average

$$\mu = \frac{15817 + 15876}{2} = 15846.5 \tag{11}$$

3. (5 points) Explain why there is not enough information to estimate how accurate the recovered data is.

   Solution For each value of θ, we have approximated the range using a polynomial of degree 1. In order to estimate the range we require information about the partial derivative $\frac{\partial^2 r}{\partial^2 w}$. There is little to no information about this derivative.

4. (5 points) Consider a target located at $d = 6000$. Estimate the elevation θ_{high} for the high trajectory to this target when $w = -2$.

   Solution Scanning the second column of the table, we see that the elevation should satisfy $\theta \in (70, 80)$ as a shell fired at $\theta = 80°$ falls short, and a shell fired at $\theta = 70°$ goes over the target. Without more information the midpoint $\theta = 75°$ is a tolerable approximation. We notice in passing that the true solution is probably much closer to $80°$ because a shot fired with this elevation falls only 482 meters short of the target, whereas a shot fired at $70°$ goes 4293 meters over the target.

5. (5 points) Estimate the relative error for θ_{high} as accurately as the data will allow.

   Solution We can say with certainty that the true elevation $\theta \in (70, 80)$. If we use the midpoint, i.e. $\theta = 75°$ as an approximation, then the relative error is bounded by $\frac{5}{70}$.

Problem 3 The trajectory of an artillery shell has been integrated numerically using time steps $h, 2h, 4h, 8h$, and $16h$. Partial results corresponding to time step h along with Richardson's fractions and error estimates are presented in Figure 3a and Figure 3b. Specifically, $x_h(t)$ is the computed approximation of the x coordinate at time t using time step h, several values of Richardson's fraction

$$F_h(t) = \frac{x_{2h}(t) - x_{4h}(t)}{x_h(t) - x_{2h}(t)}, \tag{12}$$

as well as Richardson's error estimate

$$E_h(t) = \frac{x_h(t) - x_{2h}(t)}{2^{p-1}}. \tag{13}$$

We have the same kind of information for the computed values of the y coordinate. Unfortunately, most of the data which describes the experimental setup has been lost after the numbers were produced.

Remark 1 In what follows you are asked to examine the numbers and extract evidence which supports a specific conjecture. Is it *not* enough to simply point to the relevant numbers, it is critical that you explain why the numbers support the given conjecture.

1. (6 points) Let $x(t)$ denote the shell's true x coordinate at time t and let $x_h(t)$ denote the computed approximation using time step h.

   While scanning the data for the x coordinate, what evidence do you find to support the conjecture that we have an asymptotic error expansion of the form

   $$x(t) - x_h(t) = \alpha(t)h^2 + \beta(t)h^3 + O(h^r), \quad 3 < r. \tag{14}$$

   Solution If the presence of an error expansion of the given form, Richardson's fractions $F_h(t)$ satisfy

   $$F_h(t) = \frac{x_{2h}(t) - x_{4h}(t)}{x_h(t) - x_{2h}(t)} = 2^2 \cdot \frac{1 + c_1\frac{\beta(t)}{\alpha(t)}h + O(h^{r-2})}{1 + c_2\frac{\beta(t)}{\alpha(t)}h + O(h^{r-2})} \tag{15}$$

   where $c_1 > c_2$ are positive constants. We see that $F_h(t)$ will converge monotonically to 4 and that the deviation away from 4 is $O(h)$. Scanning each row of Figure 3a, we see that the fractions converge monotonically to 4 from above and the deviation away from 4 is $O(h)$. These two observations support our conjecture.

2. (6 points) While examining the data for the x coordinate, what evidence do you find to support the conjecture that the function $\beta(t)/\alpha(t)$ does not change sign.

Solution For a given t, the behavior of Richardson's fraction is controlled by the sign of $\nu = \nu(t) = \frac{\beta(t)}{\alpha(t)}$. If $\nu > 0$, then $F_h(t)$ converges monotonically down towards 4, and if $\nu < 0$, then $F_h(t)$ converges monotonically up towards 4. Scanning the rows of Figure 3a, we only see monotonone convergence from above and down towards 4. This uniform behavior suggests that the function $\frac{\beta(t)}{\alpha(t)}$ does not change sign.

3. (3 points) What evidence do you find to support the conjecture that you have been handed data which describes a *complete* trajectory from the muzzle of the gun to the point of impact?

   Solution From the last row of the table in Figure 3b, we see that y coordinate of the shell is only about 1.33 cm. Compared with size of a standard artillery shell as well as the corresponding range which is roughly 17461 meters, this height is virtually zero and the shell will impact in an instant. This suggests that we have been given data which describes a complete trajectory.

4. (4 points) What evidence do you find to support the conjecture that the muzzle velocity was greater than 670 m/s?

   Solution The distance from the muzzle of the gun to the first data point is

   $$d = \sqrt{x_h(t_1)^2 + y_h(t_1)^2} \approx 3667m, \tag{16}$$

   where $t_1 \approx 5.5$ seconds. The shell travels along a curve path which is certainly longer than the straight line from $(0, 0)$ to $(x_h(t_1), y_h(t_1))$, so the average speed is certainly greater than $v = d/t_1 \approx 666$ m/s. Now, the muzzle veloctity is certainly larger than this average velocity, and is not a big leap of faith to suggest that is much larger than 666 m/s or even 670 m/s. In our model, air resistance is proportional to the square of the speed, so the deceleration is maximal, when the shell leaves the gun.

5. (6 points) What evidence do you find to support the conjecture that the true range is $r = 17461$ m and that this number is accurate to 5 significant figures?

   Solution Since the point $P = (x, y) = (1.74608 \cdot 10^4, 1.33293 \cdot 10^-2) \approx (17461, 0)$ belongs to the trajectory is clear that the range is close to 17461 meters. For the point P we have an error estimate which is $E_h(t) \approx 1.4744 \cdot 10^{-4}$. This error estimate is reliable, because the corresponding value of the fraction F_h is not only close to 4, but the fractions are still converging monotonically towards 4. This last observation shows that subtractive cancellation is not an issue. Since the error estimate is reliable, we deduce that the true range rounds to 17461 with five significant figures.

t	x_h(t)	F_16h(t)	F_8h(t)	F_4h(t)	F_2h(t)	F_h(t)	E_h(t)
0.000000e+00	0.00000000e+00	NaN	NaN	NaN	NaN	NaN	0.00000000e+00
5.497990e+00	3.24659494e+03	4.21151059e+00	4.10494790e+00	4.05221775e+00	4.02603872e+00	4.01300110e+00	6.31161104e-04
1.099598e+01	5.82616360e+03	4.21623805e+00	4.10629017e+00	4.05265638e+00	4.02620267e+00	4.01306956e+00	7.55420390e-04
1.649397e+01	7.97981803e+03	4.22733264e+00	4.11095146e+00	4.05477883e+00	4.02721350e+00	4.01356256e+00	7.24402208e-04
2.199196e+01	9.83700048e+03	4.24454536e+00	4.11866844e+00	4.05842372e+00	4.02898368e+00	4.01443475e+00	6.43401519e-04
2.748995e+01	1.14732693e+04	4.26881937e+00	4.12984265e+00	4.06377562e+00	4.03160160e+00	4.01572932e+00	5.48974676e-04
3.298794e+01	1.29345446e+04	4.30232889e+00	4.14552025e+00	4.07134699e+00	4.03532078e+00	4.01757239e+00	4.54824126e-04
3.848593e+01	1.42491089e+04	4.34898884e+00	4.16765184e+00	4.08210942e+00	4.04062588e+00	4.02020588e+00	3.66260487e-04
4.398392e+01	1.54345625e+04	4.41580104e+00	4.19982459e+00	4.09787408e+00	4.04842634e+00	4.02408546e+00	2.85289521e-04
4.948191e+01	1.65023604e+04	4.51632481e+00	4.24922144e+00	4.12232827e+00	4.06058912e+00	4.03015018e+00	2.12439249e-04
5.497990e+01	1.74608397e+04	4.68068524e+00	4.33257063e+00	4.16426748e+00	4.08162148e+00	4.04068146e+00	1.47447296e-04

∞ (a) Data pertaining to the x-coordinate of the shell.

t	y(h)	F_16h(t)	F_8h(t)	F_4h(t)	F_2h(t)	F_h(t)	E_h(t)
0.000000e+00	0.00000000e+00	NaN	NaN	NaN	NaN	NaN	0.00000000e+00
5.497990e+00	1.74000565e+03	4.18677009e+00	4.09294990e+00	4.04631033e+00	4.02310769e+00	4.01154122e+00	4.63609226e-04
1.099598e+01	2.85736608e+03	4.18814477e+00	4.09266937e+00	4.04595034e+00	4.02287545e+00	4.01141238e+00	5.95827378e-04
1.649397e+01	3.51854506e+03	4.19489759e+00	4.09524366e+00	4.04705027e+00	4.02338026e+00	4.01165369e+00	6.23229001e-04
2.199196e+01	3.81248313e+03	4.20542081e+00	4.09975873e+00	4.04913065e+00	4.02437735e+00	4.01214161e+00	6.12096517e-04
2.748995e+01	3.79209546e+03	4.21890015e+00	4.10576474e+00	4.05195763e+00	4.02574780e+00	4.01281619e+00	5.85319972e-04
3.298794e+01	3.49243877e+03	4.23492713e+00	4.11303874e+00	4.05541568e+00	4.02743287e+00	4.01364785e+00	5.51856330e-04
3.848593e+01	2.94013162e+03	4.25334090e+00	4.12149283e+00	4.05945906e+00	4.02940927e+00	4.01462481e+00	5.15596084e-04
4.398392e+01	2.15861212e+03	4.27421489e+00	4.13116312e+00	4.06410550e+00	4.03168577e+00	4.01575143e+00	4.78275188e-04
4.948191e+01	1.17080847e+03	4.29789306e+00	4.14222604e+00	4.06944396e+00	4.03430698e+00	4.01705008e+00	4.40577470e-04
5.497990e+01	1.33293816e-02	4.32499814e+00	4.15500228e+00	4.07563663e+00	4.03735438e+00	4.01856156e+00	4.02731126e-04

(b) Data pertaining to the y-coordinate of the shell.

Figure 3: The x and y coordinates of the shell computed using time step h. Richardson's fraction F_h is given for each value of the time and several different values of the time step. The last column contains Richardson's error estimates E_h.

Problem 4 Quake's engine hinges (among other) things on the ability to compute reciprocal square roots with an accuracy good enough to fool the human eye. Here we consider the problem of implementing a function which can compute the reciprocal square root y of any floating point number $\alpha > 0$, i.e. $x = \frac{1}{\sqrt{\alpha}}$.

1. (5 points) Explain why this problem is equivalent to solving the non-linear equation
$$f(x) = \frac{1}{x^2} - \alpha = 0, \quad x > 0. \tag{17}$$

   Solution Let $\alpha > 0$ and let $x > 0$. Then
$$x = \frac{1}{\sqrt{\alpha}} \quad \Leftrightarrow \quad x^2 = \frac{1}{\alpha} \quad \Leftrightarrow \quad \frac{1}{x^2} = \alpha \quad \Leftrightarrow \quad \frac{1}{x^2} - \alpha = 0. \tag{18}$$
   Here it is critical that $\alpha \neq 0$ and $x \neq 0$ so that we do not divide by zero at any time.

2. (5 points) Show that Newton's method for this equation can be implemented without a single division.

   Solution Evaluating f requires division, so it is perhaps a bit surprising that it is possible to avoid divisons when applying Newton's method to the problem of solving $f(x) = 0$. We have
$$\begin{aligned} x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{\frac{1}{x_n^2} - \alpha}{-\frac{2}{x_n^3}} = x_n + \frac{x_n - \alpha x_n^3}{2} \\ = x_n + \frac{1}{2} x_n (1 - \alpha x_n^2). \end{aligned} \tag{19}$$
   This last expression, i.e.
$$x_{n+1} = x_n + \left(\frac{1}{2}\right) x_n (1 - \alpha x_n^2) \tag{20}$$
   does not involve any divisions. It is necessary to multiply with the constant $\frac{1}{2}$, but this is trivial on a binary computer.

3. (8 points) Newton's method hinges on our ability to generate a good initial guess. Use your knowledge of the floating point number system to explain why it suffices to solve this problem for all $\alpha \in [1, 4]$. Obviously, we still need to handle the general case, but you need to explain why the extension from the special case is trivial in comparison with the general case.

   Solution A positive floating point number α can be written in the form $\alpha = \nu \cdot 2^m$, where the mantissa $\nu \in [1, 2)$ and the exponent m is an integer.

We have

$$\frac{1}{\sqrt{\alpha}} = \begin{cases} \frac{2^{-k}}{\sqrt{\nu}} & m = 2k \\ \frac{2^{-k}}{\sqrt{2\nu}} & m = 2k+1 \end{cases} \tag{21}$$

It follows, that we only need the ability to compute $\frac{1}{\sqrt{x}}$ where $x \in [1, 4)$. The extention to the general case is a matter of simple bit manipulations and integer arithmetic.

4. (7 points) Derive an initial guess x_0 for Newton's method such that

$$\left| x_0 - \frac{1}{\sqrt{\alpha}} \right| < \frac{1}{4} \tag{22}$$

for all $\alpha \in [1, 4]$.

Solution If $\alpha \in [1, 4]$ then $\sqrt{\alpha} \in [1, 2]$ and $\frac{1}{\sqrt{\alpha}} \in \left[\frac{1}{2}, 1\right] =: I$. It follows that the midpoint of I, i.e. $x_0 = \frac{3}{4}$ is an initial guess which is off by at most one half of the length of I, i.e. $\frac{1-\frac{1}{2}}{2} = \frac{1}{4}$.