Problem 1 Consider the problem of aiming a piece of artillery. The muzzle of the gun is located at $(0, 0)$, the target is located at $(d, 0)$ and the objective is to solve the equation

$$r(\theta) = d, \tag{1}$$

where $r(\theta)$ is the range of the gun when fired using the elevation θ. There is a single elevation ψ for which

$$r'(\psi) = 0. \tag{2}$$

The value of ψ can be treated as a known value. Let $r_{\max}$ denote the maximum range of the gun.

1. (5 points) Let $0 < d < r_{\max}$ denote the location of a target which is within range of the gun. Find a set of elevations $a < b < c$ such that

   $$a < \theta_{\text{low}} < b < \theta_{\text{high}} < c \tag{3}$$

   where θ_{low} and θ_{high} are the elevations for the low and the high trajectory from the gun to the target.

   Solution The maximum range is realized using the elevation ψ. Therefore we may choose $a = 0$, $b = \psi$ and $c = \frac{\pi}{2}$.

2. (10 points) Let a and b denote a pair of elevations, such that

   $$0 < a < b < \frac{\pi}{2} \tag{4}$$

   and

   $$r(a) < d < r(b) \tag{5}$$

   By continuity there is an elevation $\theta \in (a, b)$ such that

   $$r(\theta) = d. \tag{6}$$

   Show how to estimate both θ and the relative error in terms of a and b.

   Solution Our best estimate for the firing solution is $\theta_0 = \frac{a+b}{2}$. Since $\theta \in (a, b)$, we can say with certainty that the error satisfies $|\theta - \theta_0| \leq \frac{b-a}{2}$. Moreover, since $a < \theta$, the relative error satisfies

   $$\left| \frac{\theta - \theta_0}{\theta} \right| \leq \frac{b - a}{2a}. \tag{7}$$

   This is the best that we can do without additional information or computation.

3. (5 points) Let $0 < \bar{\theta} < \frac{\pi}{2}$ denote a good approximation of a solution θ of equation (1). Explain how to estimate the relative error regardless of the algorithm used to compute $\bar{\theta}$. Remember to explain when your estimate is reliable.

Solution We need to re-establish the situation which is present in the previous questions. We consider $a = \bar{\theta} - \Delta$ and $b = \bar{\theta} + \Delta$ where $\Delta > 0$ is a value which must carefully selected so that the following statements are true.

(a) We want $r(a) - d$ and $r(b) - d$ to have different sign, so that the interval (a, b) contains a root which can be approximated as in the previous question using $(a + b)/2 = \bar{\theta}$.

(b) We want to be able to compute the sign of $r(a) - d$ and $r(b) - d$ reliably, therefore we cannot pick Δ too small or we risk getting too close to the root. Remember, the condition number of a function is likely to have singularity near a root.

(c) Finally, we can not pick Δ arbitrarily large or we end up underestimating the quality of $\bar{\theta}$.

4. (5 points) In exact arithmetic we can always solve equation (1) exactly. In practice there is a hard limit for how accurately the equation can be solved, regardless of which algorithm is used. Explain why this is the case.

Solution Even if our computer program was able to obtain the exact result x, it would still have to round the result to the nearest floating point number $\text{fl}(x)$, which satisfies $\text{fl}(x) = x(1 + \delta)$, where $|\delta| \leq u$ and u is the unit round off error. In practice there are other concerns, such as the solutions sensitivity to rounding errors in the coefficients necessary to specify the equation, but no algorithm can obtain a better result than the floating point representation of the exact result.

Problem 2 Consider the problem of aiming a piece of artillery. The muzzle is located at $(0, 0)$ and the gun is being fired into a constant wind which blows parallel to the ground, i.e.

$$\mathbf{w} = -(w, 0), \quad w > 0. \tag{8}$$

Careful computation has established that the range of the gun is affected by the wind as follows

Elevation (degrees)	w=0 m/s Range (meters)	w=1 m/s Range (meters)	w=3 m/s Range (meters)
0	0	0	0
10	11453	11438	11410
20	15847	15817	15758
30	17670	17628	17544
40	17810	17759	17657
50	16573	16515	16400
60	14087	14026	13904
70	10415	10354	10232
80	5635	5577	5459
90	0	-56	-169

Unfortunately, the data corresponding to $w = 2$ m/s has been lost.

1. (5 points) As far as possible, do a basic sanity check of the numbers and explain why they are behaving as we have a right to expect.

   Solution For each value of the wind, the range is first an increasing function of the elevation, then it peaks around 40 degrees, and it decays afterwards. This is exactly the behavior which we have seen in class. Moreover, for each value of the elevation, the range is a decreasing function of the strength of the head wind. Again, this is extremely reasonable.

2. (5 points) Find a polynomial of degree 2 which can be used to approximate the range $r_{40,2}$ of the gun when the elevation is 40 degrees and the wind is $w = 2$ m/s.

   Solution There is exactly polynomial p of degree at most 2 such that

   $$p(w) = r_{40,w} \tag{9}$$

   for $w = 0, 1, 3$. Specifically, we have

   $$p(w) = c_0 + c_1 w + c_2 w(w - 1) \tag{10}$$

   It follows that $c_0 = r_{40,0} = 17810$. Then

   $$c_1 = \frac{r_{40,1} - c_0}{1} = 17759 - 17810 = -51. \tag{11}$$

Then

$$c_2 = \frac{r_{40,3} - c_0 - 3c_1}{3 \cdot 2} = \frac{17657 - 17810 - 3(-51)}{6} = 0 \tag{12}$$

It follows that p is really a polynomial of degree 1 and

$$p(w) = 17810 - 51w, \quad p(2) = 17708 \tag{13}$$

This polynomial has degree 1, rather than 2.

3. (5 points) Given the above information, what is the best error bound that you have for the range $r_{40,2}$ and what additional information do you need in order to compute a better bound?

   Solution We can say with certainty that the range $r_{40,2}$ is between $r_{40,1} = 17759$ and $r_{40,3} = 17657$. The average is $r_{40,2} \approx 17708$. The corresponding error bound is $\frac{r_{40,3} - r_{40,1}}{2} = 51$. We need information about the size of the derivatives of r with respect w in order to obtain a better error bound. In order to estimate the quality of the approximation $r_{40,2} \approx p(2)$ we need an upper bound on $\frac{\partial^3 r}{\partial^3 w}$.

4. (5 points) Give a physical reason why the case of $w = 1000$ m/s is absurd and irrelevant.

   Solution On our planet, maximum wind speeds are less than 200 m/s. The study of wind speeds larger than those encountered on Earth are not particularly relevant until the time where a FLT drive is implemented.

5. (5 points) Give a purely mathematical reason why the case of $w = 1000$ m/s can not be investigated using the data given.

   Solution The error formula for polynomial interpolation does not ensure that the error is small when we extrapolate far outside the interval which contains the nodes. In practice, this ensures that the error will not be small.

Problem 3 The trajectory of a shell has been been computed numerically using one of the standard Runge-Kutta methods and time steps $2^j h$, where $h > 0$ and $j = 0, 1, 2, 3, 4, 5, 6$. The computed x coordinates corresponding to the smallest time step, as well as several auxiliary numbers are given in Figure 1. These numbers include Richardson's fractions

$$F_h = \frac{x_{2h} - x_{4h}}{x_h - x_{2h}}, \tag{14}$$

and his error estimates

$$E_h = \frac{x_h - x_{2h}}{2^p - 1}. \tag{15}$$

Consider the hypothesis that there exists functions α and β such that

$$x(t) - x_h(t) = \alpha(t)h^p + \beta(t)h^q + O(h^r), \quad 0 < p < q < r, \tag{16}$$

where $x(t)$ is the x coordinate of the shell at time t and $x_h(t)$ is the computed value of the x coordinate of the shell at time t based on the time step.

1. (5 points) Scanning the rows of Figure 1 what evidence can you find to support this hypothesis?

   Solution For each fixed value of the time, i.e. for each row, we see that Richardson's fractions are converging monotonically towards 4 as the time step decreases.The convergence is either from above (for $t \leq 0.6$) or from below (for $t > 0.6$).

2. (5 points) What evidence do you find to support the idea that $p = 2$?

   Solution If the hypothesis is correct, then Richardson's fraction will converge to 2^p. Since we are using a Runge-Kutta method and since all functions involved are smooth, we are certain that p is an integer. Evidently $p = 2$ is the only choice.

3. (5 points) What evidence do you find to support the idea that $q = 3$? The difference between q and p determines the speed at which the Richardson's fraction converge towards 2^p. Specifically $F_h = 2^p + O(h^{q-p})$. As the time step is reduced by a factor of 2, it is evident that the deviation away from $2^2 = 4$ is also reduced by a factor of 2. Evidently, $q - p = 1$ or $q = 3$.

   Solution

4. (5 points) What evidence do you find to support the idea that $\frac{\beta(t)}{\alpha(t)}$ changes sign in the interval from 60 to 70 seconds?

   Solution Richardson's fraction for a particular value of the time t will (ultimately) converge monotonically up (down) towards 2^p if $\frac{\beta(t)}{\alpha(t)} < 0$ ($\frac{\beta(t)}{\alpha(t)} > 0$). Evidently, the function $\frac{\beta(t)}{\alpha(t)}$ must change sign somewhere in the interval $60, 70$).

5. (5 points) Given that the kill radius of the shell is 15 m what evidence do you find to support the idea that the time step h is far smaller than absolutely necessary in order to compute an adequate firing solution?

   Solution Since each row of fractions are converging monotonically towards 4 and since each value of $F(h)$ is close to 4 there is no reason to doubt the magnitude, sign and the first couple of digits of each error estimate. These are all in the range of millimeters! This level of accuracy is ridiculous given that the kill radius of the shell is 15 m.

   Remark 1 Pinpoint accuracy is required in the context of Close-In Weapons Systems (CIWS) which defend a ship against sea-skimming missiles. Here the effective range is a few kilometers. The missiles are travelling much faster than the speed of sound and have to be destroyed at least 500 meters or shrapnel from the explosion can still kill or injure unshielded personal or damage antenna. Missiles have a small cross section and denotating the warhead requires centimeter accuracy.

t (second	x(h) (meters)	F(16h)	F(8h)	F(4h)	F(2h)	F(1h)	E(h) (meters)
0.000000e+00	0.00000000e+00	NaN	NaN	NaN	NaN	NaN	0.00000000e+00
1.000000e+01	4.42636428e+03	4.40208302e+00	4.19652547e+00	4.09686835e+00	4.04805949e+00	4.02393324e+00	2.02982097e-03
2.000000e+01	7.59249294e+03	4.45743911e+00	4.21826652e+00	4.10638756e+00	4.05249823e+00	4.02607450e+00	1.81580364e-03
3.000000e+01	1.01303686e+04	4.55971288e+00	4.26316429e+00	4.12731515e+00	4.06258854e+00	4.03102724e+00	1.33200320e-03
4.000000e+01	1.22788706e+04	4.75684747e+00	4.35446129e+00	4.17104781e+00	4.08396761e+00	4.04159417e+00	8.52917015e-04
5.000000e+01	1.41255017e+04	5.22603846e+00	4.58941697e+00	4.28826081e+00	4.14246662e+00	4.07081183e+00	4.30192140e-04
6.000000e+01	1.56951869e+04	7.26961082e+00	5.99834507e+00	5.15708498e+00	4.63447539e+00	4.33433302e+00	7.88617763e-05
7.000000e+01	1.70007109e+04	-1.55835424e+00	2.57912803e+00	3.45735493e+00	3.76068409e+00	3.88742139e+00	-2.05576781e-04
8.000000e+01	1.80630173e+04	2.41140693e+00	3.46805668e+00	3.78135133e+00	3.90087282e+00	3.95280986e+00	-4.37688787e-04
9.000000e+01	1.89121966e+04	3.03968536e+00	3.66792881e+00	3.86268274e+00	3.93770491e+00	3.97035220e+00	-6.31441520e-04
1.000000e+02	1.95822498e+04	3.29433745e+00	3.75543545e+00	3.89934326e+00	3.95452039e+00	3.97840897e+00	-7.96168201e-04

Figure 1: The partial results of tracking an artillery shell numerically for $T = 100$ seconds using a standard Runge-Kutta method and time steps $2^j h$, where the smallest time step is $h = 2^{-5}$ seconds. The x-coordinate of the shell is provided every 10 seconds. Richardson's fractions and his standard error estimate has been computed for each of these times.

Problem 4 Consider the problem of computing the side a in an arbitrary triangle ABC, see figure 2. In exact arithmetic we have the formula

$$a^2 = b^2 + c^2 - 2bc\cos(A) \tag{17}$$

where a, b, and c are sides and A is the angle opposing the side A.

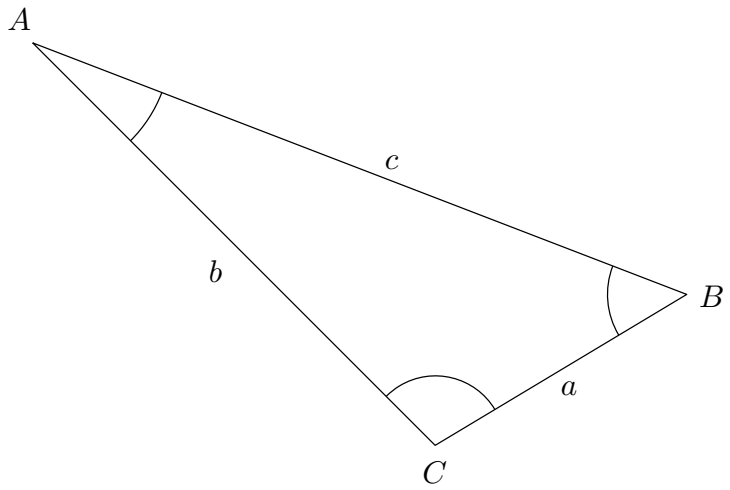

Figure 2: A general triangle ABC with sides a, b, and c.

1. (5 points) Characterize the set of triangles for which this formula is likely to fail miserably when the computations are carried out in floating point arithmetic.

   Solution Assuming that there exists a reliable implementation of the function $\theta \to \cos(\theta)$, then the only potential problem is the single subtraction, which can cancel catastrophically. This happens precisely, then A is tiny and $b \approx a$. These triangles are needle shaped. In order to *prove* that this characterization is valid we consider the extreme case, i.e. equation

   $$b^2 + c^2 = 2bc\cos(A). \tag{18}$$

   It follows, that

   $$\begin{aligned} 0 \leq (b-c)^2 = b^2 + c^2 - 2bc \\ = 2bc\cos(A) - 2bc = 2bc(\cos(A) - 1) \leq 0. \end{aligned} \tag{19}$$

   which is only true if $\cos(A) = 1$ or equivalently $A = 0$. Since we now know that

   $$b^2 + c^2 = 2bc \tag{20}$$

it follows immediately that $(b-c)^2 = 0$ or $b = c$. In floating point arithmetic it is the near occurrence of this situation which is problematic.

Remark 2 It does not matter if b, c and A are machine numbers. In all likely-hood neither $a^2 + b^2$, nor $\cos(A)$ are *not* a machine number and the subtraction will then emphasize the rounding error associated with these numbers. There is no guarantee that the result will have even the right sign!

2. (10 points) Show that the formula

$$a^2 = (b-c)^2 + 4bc\sin^2\left(\frac{A}{2}\right). \tag{21}$$

is mathematically equivalent to formula (17).

Hint: Trigonometric identities are a pain to memorize. Fortunately, the problem can be completed using the definitions

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}, \tag{22}$$

where i is the imaginary unit, $i^2 = -1$.

Solution By comparing the two expressions we see that it suffices to show that

$$b^2 + c^2 - 2bc + 4bc\sin^2\left(\frac{A}{2}\right) = b^2 + c^2 - 2bc\cos(A) \tag{23}$$

or equivalently

$$4\sin^2\left(\frac{A}{2}\right) = 2(1-\cos(A)). \tag{24}$$

Starting with the left hand side, we find that

$$\begin{aligned} 4\sin^2\left(\frac{A}{2}\right) = 4\left(\frac{e^{iA/2} - e^{-iA/2}}{2i}\right)^2 &= -\left(e^{iA} + e{-}iA - 2\right) \\ &= \left(2 - 2\frac{e^{iA} + e^{-iA}}{2}\right) = 2(1-\cos(A)), \end{aligned} \tag{25}$$

which is exactly the right hand side of equation (24).

3. (10 points) Why will equation (21) always provide sensible results even in floating point arithmetic?

Hint: The real goal is to compute $a > 0$ rather than a^2.

Solution Even if a, b, and A are floating point numbers, there is no guarantee that $a^2 + b^2$ and $\cos(A)$ are floating point numbers. If $a \approx b$

and $A \approx 0$, then the subtraction $d = (a^2 + b^2) - (2bc\cos(A)$ will cancel catastrophically. In particular, we can not be certain that the compute value $\hat{d}$ has the correct sign and $\hat{d} < 0$ is a distinct possibility. Since the real goal is compute $a = \sqrt{a^2}$ the square root operation will either fail with or without a warning or return a purely imaginary number, and the effects of this spurious result on the rest of the program are unknown. In contrast, the equivalent expression given by equation (21) does not suffer from this problem, because all intermediate results are guaranteed to be non-negative. In particular the computed argument for the square root function will be non-negative and if the square root function has been implemented correctly, then it will not fail.

Remark 3 However, the subtraction $(b-c)$ included in equation (21) will still cancel catastrophically if the real numbers b and c satisfy $b \approx c$. If there are formulas for b and c which the permit the expression $b - c$ to be rewritten, then this option should certainly be explored, otherwise the standard recourse is compute b and c as well as $b - c$ in higher precision. However, if neither option is feasible, and if $b \approx c$ and $A \approx 0$, then the right hand side of equation (21) will be a small positive number with a large relative error. This *can* be a disaster if you are dealing with a large number of triangles which are all needle-shaped and participte in an unfortunate manner in larger calculation. A specific example is computing the area of country by triangulation. In this case, the problem evaporates and you get a reliable final result for the area of the country, if there is just a single large triangle with a decent shape. If the original formula is applied to a large number of needles, then you risk ending up in the Land of Make Believe with a final result which is complex!