Problem 1 Consider the problem of computing the function $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) = \frac{\sqrt{1 + \sin(x)^2} - 1}{x}, \quad x > 0 \tag{1}$$

1. Show that $f(x) > 0$ for all $x > 0$.

   Solution There was a mistake in this problem, because $f(k \cdot \pi) = 0$ for all integer k. However, for $x \in (0, \pi)$, we have $\sin(x) > 1$ from which it follows that $\sqrt{1 + \sin(x)^2} > 1$ and $f(x) > 0$. In general, we have $f(x) \geq 0$ and $f(x) = 0$ if and only if $x = k\pi$ for some integer k.

2. Show that $f(x) \to 0$ for $x \to 0_+$.

   Solution There is more than one way to solve this problem.

   (a) Using l'Hospital's rule is one option. We have

   $$f(x) = \frac{T(x)}{N(x)}, \tag{2}$$

   where

   $$\begin{aligned} T(x) &= \left(1 + \sin(x)^2\right)^{\frac{1}{2}} - 1 \to 0, & x &\to 0_+, & (3) \\ N(x) &= x \to 0, & x &\to 0_+. & (4) \end{aligned}$$

   We therefore examine the behavior of T' and N'. We have

   $$\begin{aligned} T'(x) &= \left(1 + \sin(x)^2\right)^{-\frac{1}{2}} \sin(x) \cos(x) \to 0, & x &\to 0_+, & (5) \\ N'(x) &= 1 \to 1, & x &\to 0_+. & (6) \end{aligned}$$

   By l'Hospital's rule, it follows that

   $$f(x) \to 0, \quad x \to 0_+. \tag{7}$$

   (b) A faster alternative is to rewrite f as

   $$f(x) = \frac{\sin(x)^2}{x(\sqrt{1 + \sin(x)^2} + 1)} \tag{8}$$

   Since $\frac{\sin(x)}{x} \to 1$ as $x \to 0$, it follows that $f(x) \to 0$ as $x \to 0$.

3. Figure 1 shows the results of the MATLAB commands

```
a=0; b=2e-7;
s=linspace(a,b,1025);
f=@(x)(sqrt(1+sin(x).^2)-1)./x;
plot(s,f(s),'LineWidth',2); grid on; grid minor;
xlabel('x'); ylabel('y=f(x)');
```

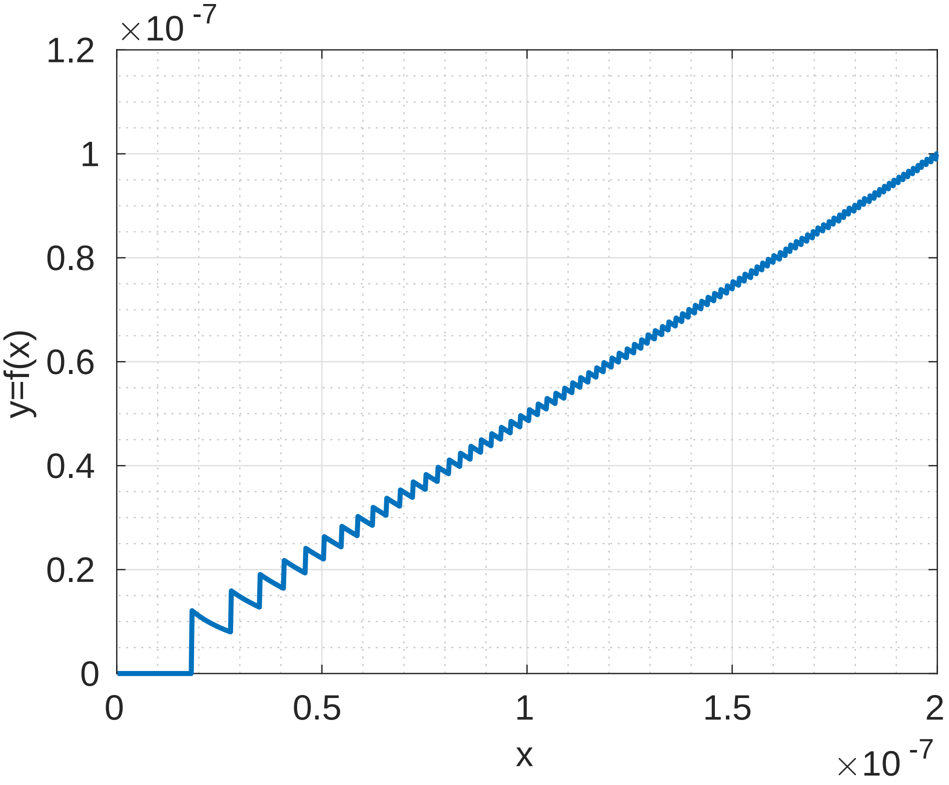

Figure 1: The result of a naive computation of $f$using MATLAB

Why is it immediately clear that this is not a numerically reliable way of computing f? Give as many reasons as you can!

Solution

(a) We know that $f(x) > 0$ for $x \in (0, \pi)$, yet the naive computation appears to return 0 for all x in an interval close to 0.

(b) It is clear from the definition that f will exhibit damped oscillations, but the period will be controlled by the period of $x \to \sin(x)$, i.e. it will oscillate much more slowly than the graph indicates.

(c) The computed rapid oscillations are not typical of smooth function such as f. They are however typical of catastrophic cancellation.

4. Explain why $\psi(x) = \frac{x}{2}$ is a good approximation of f for $x \in [0, 2 \times 10^{-7}]$?

Solution This is a question of applying Taylor's theorem. This can be done in more than one way.

(a) We have

$$\sin(x) \approx x, \quad \sqrt{1+x} \approx 1 + \frac{1}{2}x \tag{9}$$

for small values of x. It follows that

$$f(x) \approx \frac{\sqrt{1+x^2}-1}{x} \approx \frac{(1+\frac{1}{2}x^2)-1}{x} = \frac{1}{2}x \tag{10}$$

(b) It is also possible to write

$$T(x) = T(0)+T'(0)x+\frac{1}{2}T''(0)x^2+O(x^3) = \frac{1}{2}T''(0)x^2+O(x^3) \tag{11}$$

because $T(0) = T'(0) = 0$. It is important to realize that we do not need the complete expression for $T''(x)$! From

$$T'(x) = \left[\left(1+\sin(x)^2\right)^{-\frac{1}{2}}\cos(x)\right]\sin(x) \tag{12}$$

we see that

$$\begin{aligned} T''(x) = {} & \left[\left(1+\sin(x)^2\right)^{-\frac{1}{2}}\cos(x)\right]'\sin(x) \\ & + \left[\left(1+\sin(x)^2\right)^{-\frac{1}{2}}\cos(x)\right]\cos(x) \end{aligned} \tag{13}$$

It follows that $T''(0) = 1$ and

$$f(x) \approx \frac{\frac{1}{2}x^2}{x} = \frac{1}{2}x \tag{14}$$

(c) Direct computation of $f'(0)$ from the definition is also possible of a derivative is also possible. We have $f(0) = 0$, so

$$\frac{f(x)-f(x)}{x-0} = \frac{\sqrt{1+\sin(x)^2}-1}{x^2}. \tag{15}$$

This is also a form which suggests an application of l'Hospital's rule. A short calculation which capitalizes on the above work on T shows that

$$f'(0) = \frac{1}{2} \tag{16}$$

It follows, that

$$f(x) \approx \frac{1}{2}x. \tag{17}$$

As for order of the approximation it is clear from the definition that f is an odd function. It follows that

$$f(x) = \frac{1}{2}x + O(x^3) \tag{18}$$

5. What is the largest relative error associated with the naive approximation of $f(x)$, when $x \in [0, b]$, where $b = 2 \times 10^{-7}$?

   Solution It is clear that the amplitude of the oscillations away from the excellent approximation $y = \frac{1}{2}x$ are at all times vastly smaller than $\frac{1}{2}x$ except for $0 < x \leq \beta$ where $\beta \approx 0.2 \times 10^{-7}$. Thus the relative error is small and certainly less than 1 on the interval (β, b). Inside the interval $(0, \beta)$ the relative error is 1. It follows that largest relative error is 1.

Problem 2 Consider the problem of solving a non-linear equation

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

where $g : \mathbb{R} \to \mathbb{R}$.

1. Let $a \neq b$ and suppose $g(a)g(b) < 0$. Which property of g will allow you to draw the nontrivial conclusion that g has at least one root between a and b?

Solution By assumption $g(a)$ and $g(b)$ have distinct signs and they are both nonzero. If g is also continuous, then by the intermediate value theorem, there exists at least one c between a and b such that $g(c) = 0$. The critical property is continuity of g.

Remark 1 We notice in passing that the function g given by

$$g(x) = \begin{cases} -1 & x < 0, \\ c & x = 0, \\ 1 & x > 0, \end{cases} \tag{20}$$

satisfies $g(-1)g(1) < 0$, but unless $c = 0$ there is no zero in $(-1, 1)$ or anywhere else for that matter. Moreover, regardless of the value of c, g will be discontinous at the point $x = 0$.

Now consider an iterative method for solving our non-linear equation (19). Let x_j denote the exact value of the jth approximation, and let $\hat{x}_j$ denote the computed value of x_j.

2. Assume that the *computed* residual $\hat{g}(\hat{x}_j)$ is exactly equal to 0 for some value of j. Explain, why this should not be taken as evidence that $\hat{x}_j$ is a root of g.

   Solution Rounding errors are essentially unavoidable, hence $\hat{g}(\hat{x}_j) \neq g(\hat{x}_j)$ *unless* we are exceptionally lucky. It follows that $\hat{g}(\hat{x}_j) = 0$ does not imply that $\hat{x}_j$ is a root of g.

3. Give an example of an iterative method, a function g and an initial guess x_0 for a root r for which $|g(x_j)| < \epsilon$, where ϵ is a tiny number, but x_j is far from any root of g for all large values of j.

   Solution It suffices to sketch the graph of such a function. The critical point to get right is the asymptotic behavior of g. An explicit example is

   $$g(x) = \left(1 - \frac{1}{x}\right) e^{-\lambda x}, \quad \lambda > 0, \tag{21}$$

   for which

   $$g'(x) = \frac{1}{x^2} e^{-\lambda x} - \lambda \left(1 - \frac{1}{x}\right) e^{-\lambda x} \tag{22}$$

so that Newton's iteration takes the form

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} = x_n - \frac{1 - \frac{1}{x_n}}{\frac{1}{x_n^2} - \lambda\left(1 - \frac{1}{x_n}\right)}$$

$$= x_n - \frac{x_n^2 - x_n}{1 - \lambda(x_n^2 - x_n)} \tag{23}$$

In particular, if x_n is large, then

$$x_{n+1} \approx x_n + \frac{1}{\lambda} \tag{24}$$

which show that

$$x_n \to \infty, \tag{25}$$

provided that x_0 is chosen sufficiently large. It follows, that $g(x_n) \to 0$, while x_n moves away from the only root $r = 1$.

4. What are the advantage(s)/disadvantage(s) of the bisection method?

   Solution

   (a) The advantage of the bisection method is that it is robust. It maintains a bracket around the root which is repeatedly cut in two until convergence.

   (b) Moreover, we only need to calculate the sign of g, so we only need to approximate g with a relative error which is strictly less than 1.

   (c) The disadvantage is that the convergence is very slow.

   (d) It might be problematic to compute g with a small relative error when we are close to the root, but this issue affects all methods for solving $g(x) = 0$.

5. Describe an iterative method which is both fast and robust.

   Solution Such an example is afford by the robust secant method. It maintains a bracket around the root which ensures that it is robust. If a secant step is possible, then it refines the bracket using the new point. If a secant step is impossible (division by zero), then the bracket is refined using bisection. In practice we often see a few initial bisection steps followed by a short sequence of secant steps which converge rapidly. The result is a method which is as robust as the bisection method and nearly as fast as the secant method.

Problem 3 The integral $I = \int_0^1 \phi(x)dx$ of an unknown function $\phi : [0,1] \rightarrow \mathbb{R}$ has been computed numerically using the trapezoidal rule and subjected to Richardson's techniques. Figure 3 contains all the available data. The number N is the number of sub-intervals, A_h is the computed approximation corresponding to the step size $h = 1/N$. Richardson's fraction is the number

$$F_h = \frac{A_{2h} - A_{4h}}{A_h - A_{2h}} \tag{26}$$

As there is some doubt about the order p of the method, the numbers $A_h - A_{2h}$ has been provided instead of the usual error estimates given by

$$E_h = \frac{A_h - A_{2h}}{2^p - 1}. \tag{27}$$

1. What evidence do you find to support the conjecture that ϕ is not infinitely often differentiable?

   Solution If ϕ was infinitely often differentiable, then Richardson's fraction F_h would converge to $2^2 = 4$ as h tends to zero. This is clearly not the case, so ϕ is not infinitely often differentiable on $[0,1]$.

2. What is the order p of the dominant term of the asymptotic error expansion (AEX) for the trapezoidal method when applied to ϕ?

   Solution If an AEX of the form

   $$T - A_h = \alpha h^p + \beta h^q + O(h^r), \quad 0 < p < q < r \tag{28}$$

   exists, then

   $$F_h \rightarrow 2^p, \quad h \rightarrow 0 \tag{29}$$

   From the table we find that $p = 1.5$ is likely, because $2^{1.5} \approx 2.828$. In principle, any value of p is possible, but in the past, we have only encountered nice rational values of p which is why we select $p = 1.5$ as the most likely order.

3. What evidence do you find to support the conjecture that the approximations continue to improve as long as $N \leq 2097152$?

   Solution As long as $N \leq 2097152$ we see Richardson's fraction approach $2^{1.5}$, the approach is even monotone from below for all but the first few values. For $N = 4194304$ Richardson's fraction has jumped to the other side of $2^{1.5}$. This is evidence that the subtrative cancellation is setting in. In the past, we have seen that this is also the point where the error estimate starts to loose accuracy.

4. What are all the consequences of increasing the number of sub-intervals beyond this point?

   Solution There are numerous points to be made here.

(a) Moving down one row in the table doubles the number of subintervals and function eveluations needed. If ϕ is costly to compute, then this is something which we want to avoid.

(b) As we increase N, the approximations $A(h)$ converge to the target value T. It follows that the computed approximations $\hat{A}_h$ are all close to the same real number, i.e. T. Therefore, the subtractions $\hat{A}_h - \hat{A}_{2h}$ will cancel catastrophically.

(c) As we increase N, then we will gradually reach a point where the rounding errors dominate and the calculations are rendered meaningless.

(d) In particular, the computed value of Richardson's fraction will deviate significantly from the theoretical behavior, i.e. monotone convergence towards 2^p and the error estimates will gradually become less accurate even to the point where the sign can be wrong.

5. What is the smallest value of N which will allow you to compute the integral I with a relative error less than $\tau = 10^{-6}$?

Solution We want the error to satisfy $|E| < \tau|T|$, which is why we seek an *trustworthy* error estimate E_h

$$|E_h| < \tau|T|. \tag{30}$$

or equivalently

$$|A_h - A_{2h}| < (2^p - 1)\tau|T|, \quad p = \frac{3}{2} \tag{31}$$

It is apparent that $T \approx 1.25$ is not a bad approximation and that $1.25 < T$. Therefore we seek N such

$$|A_h - A_{2h}| < (2^{1.5} - 1) \times 10^{-6} \times 1.25 \approx 2.74 \times 10^{-6} \tag{32}$$

This happens the first time in when $N = 2^{12} = 4096$. We stress that this value of N is well inside the range of values for which Richardson's fraction is converging towards $2^{1.5}$ in a monotone manner. This indicates that the higher order terms of the AEX are insignificant and that the error estimate is reliable.

```
      N |       Approximation A(h) |    Richardson's fraction |   A(h)-A(2h)
      1 |   1.3591409142295228e+00 |   0.0000000000000000e+00 |   0.0000000000000000e+00
      2 |   1.2624814525140426e+00 |   0.0000000000000000e+00 |  -9.6659461715480122e-02
      4 |   1.2500878518926526e+00 |   7.7991428535026399e+00 |  -1.2393600621390055e-02
      8 |   1.2516121251639316e+00 |  -8.1308259187609906e+00 |   1.5242732712790197e-03
     16 |   1.2536843987528026e+00 |   7.3555599968317309e-01 |   2.0722735888709654e-03
     32 |   1.2548090999163359e+00 |   1.8425103983717011e+00 |   1.1247011635333592e-03
     64 |   1.2553062339580381e+00 |   2.2623700434644585e+00 |   4.9713404170215192e-04
    128 |   1.2555071315593258e+00 |   2.4745643477851758e+00 |   2.0089760128771950e-04
    256 |   1.2555844894995216e+00 |   2.5969874686322156e+00 |   7.7357940195810215e-05
    512 |   1.2556134303720221e+00 |   2.6729650322218466e+00 |   2.8940872500493597e-05
   1024 |   1.2556240616527965e+00 |   2.7222376226031502e+00 |   1.0631280774386909e-05
   2048 |   1.2556279204209184e+00 |   2.7550970772221004e+00 |   3.8587681219226511e-06
   4096 |   1.2556293097578939e+00 |   2.7774169910282742e+00 |   1.3893369754658380e-06
   8192 |   1.2556298072348882e+00 |   2.7927662816183387e+00 |   4.9747699426561098e-07
  16384 |   1.2556299846890373e+00 |   2.8034114538352468e+00 |   1.7745414915282254e-07
  32768 |   1.2556300478211850e+00 |   2.8108365631201058e+00 |   6.3132147731792543e-08
  65536 |   1.2556300702399685e+00 |   2.8160380665719309e+00 |   2.2418783496291894e-08
 131072 |   1.2556300781907634e+00 |   2.8196908291928016e+00 |   7.9507949113377663e-09
 262144 |   1.2556300810079308e+00 |   2.8222657072051205e+00 |   2.8171673882582127e-09
 524288 |   1.2556300820055022e+00 |   2.8240259338450122e+00 |   9.9757135885170101e-10
1048576 |   1.2556300823585662e+00 |   2.8254687404681209e+00 |   3.5306402246249036e-10
2097152 |   1.2556300824834798e+00 |   2.8264650411244290e+00 |   1.2491363499123054e-10
4194304 |   1.2556300825276043e+00 |   2.8309371524615159e+00 |   4.4124481846097297e-11
8388608 |   1.2556300825432214e+00 |   2.8254020161232991e+00 |   1.5617063198192227e-11
```

Figure 2: The results of integrating ϕ numerically using the trapezoidal rule and a selection of step-sizes

Problem 4

An artillery trajectory has been computed. The script used to compute all the numbers is given in Figure 3. A plot of the computed trajectory is given in Figure 4. Auxiliary data computed by the script is given in Figure 5 and Figure 6.

1. What evidence do you find to support the conjecture that this is a flat trajectory?

   Solution We should not attach too much importance to the fact that $\theta = 20\frac{\pi}{180}$ is specified directly in the script, after all, the main files are not available, and they might contain an error. However, by examining Figure 4 we see that the apex point is less than 2 km, while the range is certainly beyond 13 km, observations which are consistent with a flat trajectory.

   Remark 2 It became apparent during the grading process that I had not used the phrase "flat trajectory" during the lectures. It is a commonly used term in ballistics literature and refers to a low elevation. In this case it is possible compute approximate firing solutions much more rapidly. The grading scale was adjusted to reflect this oversight on my part.

2. What evidence do you find to support the conjecture that air resistance was included in the calculation?

   (a) An implementation of the international standard atmosphere model is explicitly passed to the computation. However, this does not prove conclusively that it is used correctly!

   (b) In the abscence of air resistance, the trajectory would be symmetric around the line $x = x_a$ where (x_a, y_a) is the apex point. The trajectory is not symmetric, in fact $x_a \approx 8$ km, while the range is less than 14 km, indicating that the shell is slowing down. This is consistent with the precense of friction.

3. What evidence do you find to support the conjecture that the shell impacted the ground with a speed which was less than 297 meters/s.

   The script computes the terminal state of the projectile as a function of the time step size h. The terminal velocity is compute in the first line of the second to last block. Richardson's techiques are applied to the terminal velocity in the first line of the last block, resulting in the first table, i.e. Figure 5. Examining this table, we see Richardson's fraction converging to 4 in a manner which is consistent with an AEX of the form

   $$T - A_h = \alpha h^2 + \beta h^3 + O(h^r), \quad 3 < h^r \tag{33}$$

   Admittedly, the table is rather short, and most people whould prefer a few more rows. However, and this is critical, the error estimates are negative,

```
% Load preprogrammed drag coefficients
load shells
% Define the shell and enviroment
param.mass=10;
param.cali=0.088;
param.drag=@(x)mcg7(x);
param.atmo=@(x)atmosisa(x);
param.grav=@(x)9.82;
% Define shot
v0=830; theta=20*pi/180;

% Various other stuff.
method='rk2'; dt=1; maxstep=200;
[r, flag, t, tra]=range_rkx(param,v0,theta,method,dt,maxstep);
% Generate plot.
plot(tra(1,:),tra(2,:),'k-','LineWidth',4); grid on; grid minor;
xlabel('x (meters)'); ylabel('y (meters)');

% Allocate space for data collection
data=zeros(4,5);
% Data collection
for k=1:5
  [r, flag, t, tra]=range_rkx(param,v0,theta,method,dt,maxstep);
  data(:,k)=tra(:,end);
  dt=dt/2; maxstep=maxstep*2;
end

% Compute speed of impact
v=sqrt(data(3,:).^2+data(4,:).^2);
% Compute some angle
angle=atan(data(3,:)./data(4,:));

% Run Richardson's scheme, generates the first table
richardson(v,2);
% Run Richardson's scheme, generates the second table
richardson(angle,2);
```

Figure 3: The script used to compute the trajectory and all auxiliary numbers.

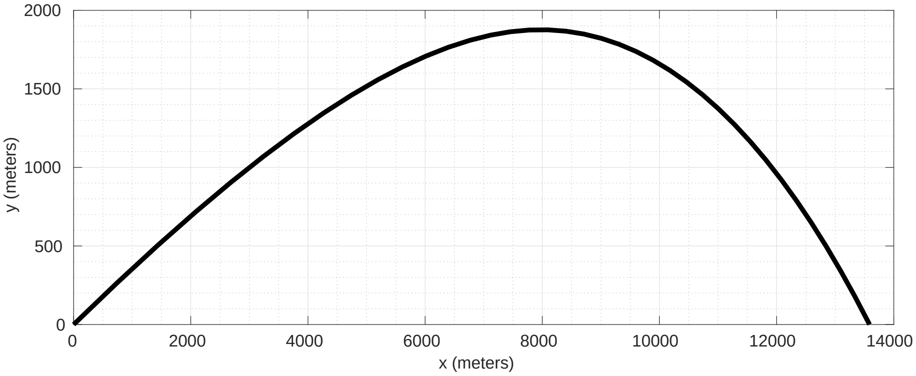

Figure 4: The computed trajectory of an artillery shell.

```
k |        approximation |     fraction |       error estimate
1 |   2.961508712574e+02 |   0.00000000 |   0.000000000000e+00
2 |   2.961018547930e+02 |   0.00000000 |  -1.633882148523e-02
3 |   2.960888812097e+02 |   3.77817473 |  -4.324527759119e-03
4 |   2.960859000494e+02 |   4.35185705 |  -9.937200846745e-04
5 |   2.960851842022e+02 |   4.16452023 |  -2.386157417125e-04
```

Figure 5: The first table generated by the script

```
k |        approximation |     fraction |       error estimate
1 |  -9.374500477428e-01 |   0.00000000 |   0.000000000000e+00
2 |  -9.368320178834e-01 |   0.00000000 |   2.060099531333e-04
3 |  -9.366844734718e-01 |   4.18877172 |   4.918147054445e-05
4 |  -9.366488584576e-01 |   4.14275875 |   1.187167139971e-05
5 |  -9.366401377402e-01 |   4.08395461 |   2.906905816465e-06
```

Figure 6: The second table generated by the script

which indicate that the computed terminal velocities are too large! Since they are all about 296 m/s, it is clear that the true terminal velocity is certainly less than 297 m/s.

Remark 3 The muzzle velocity was specified as 830 m/s. Since the muzzle and the point of impact have the same height, it is apparent that the shell has lost a significant amount of mechanical energy. This provides an additional and very strong indication that the simulation too air resistance into account.

4. What evidence do you find to support the conjecture that the shell did not impact the ground with an angle of 45 degrees?

   Solution The angle of impact ψ is not defined explicitly in the text. The only sensible choices are the angle between the velocity and the x-axis or the y-axis. It is possible to toggle the sign as well so that there are 4 distinct definitions. They are however closely related. The script settles the issue, by choosing

$$\psi = \tan^{-1}\left(\frac{v_x}{v_y}\right) \tag{34}$$

   and it is up to the user to make geometrical sense of this number. However, and this is critical, the numerical value is found to be roughly 9.3664×10^{-1} radian or equivalently 53.6655 degrees. We are not confident that the error estimate has many accurate digits, but certainly there is no reason to suspect that the order of magnitude ($O(10^{-6})$ radian) is wrong. Hence it is unlike in the extreme that the true angle of impact is 45 degrees.

5. Why is it good physics that the speed of the shell at the point of impact is smaller than the muzzle velocity?

   Solution The shell leave the muzzle at $(0, 0)$ with speed 830 m/s and impacts the ground at $(r, 0)$ with terminal velocity less than 297 m/s. The total mechanical energy has dropped significantly. This is consistent with the atmosphere doing work on the shell. Any program claiming to simulate real physics should certainly show a drop in mechanical energy.