

Problem 1 Table 1 contains all available information about a specific polynomial $p : \mathbb{R} \rightarrow \mathbb{R}$. The computed value \hat{y} of $y = p(x)$ as well as a running error bound $\mu = \mu(x)$ is given for many different values of x . The running error bound satisfies

$$|y - \hat{y}| \leq \mu u, \quad (1)$$

where u is the double precision unit round off, $u = 2^{-53}$.

1. (5pt) Why can you immediately deduce that p has a root x_2 in the vicinity of $x = 2$?

Solution We can trust the computed sign of p at the points $x = 1.98$ and $x = 2.02$. Moreover, $p(x)$ changes sign from $x = 1.98$ to $x = 2.02$. It follows, that p must have a zero between these two points. This is true because p is continuous because it is a polynomial.

2. (5pt) Why is not trivial to determine if p has a root x_1 in the vicinity of $x = 1$?

Solution There is no change in the computed value of the sign. The fact that the computed value \hat{y} of $y = p(1)$ equals 0 is not evidence of the fact that the true value satisfies $y = 0$. It is entirely possible, that the zero is the effect of a rounding error.

3. (5pt) Compute the root x_2 as accurately as the data will allow.

Solution By question 2 we have $x_2 \in (1.98, 2.02)$. In the absence of any additional information of best approximation is $a_2 = 2.0$. It follows that $|x - a_2| \leq 0.02$. Moreover, since $1.98 < x_2$, we have

$$\frac{|x_2 - a_2|}{|x_2|} < \frac{0.02}{1.98} = \frac{1}{99} \quad (2)$$

4. (5pt) Why can you be certain that p' has a root in the vicinity of $x = 1$?

Solution The running error bounds are so small, that there is no doubt that $p(0.96) < p(0.98)$ and $p(1.02) > p(1.04)$. By the mean value theorem, it follows that $p'(x)$ is positive for at least one $x \in (0.96, 0.98)$ and $p'(x)$ is negative for at least one $(1.02, 1.04)$. It follows by the continuity of p' that there exists an $x \in (0.96, 1.04)$ such that $p'(x) = 0$.

5. (5pt) What evidence do you find to support the conjecture that p has a double root at $x = 1$.

If p has a double root at $x = 1$, then $p(x) = (x - 1)^2 q(x)$ where $q(1) \neq 0$. Let $\tilde{p}(x) = p(x - 1) = x^2 q(x - 1)$. For small values of x we have $\tilde{p}(x) \approx x^2 q(0)$ which implies

$$g(x) = \frac{\tilde{p}(2x)}{\tilde{p}(x)} \approx 4 \quad (3)$$

x	y	mu		x	y	mu
8.2000e-01	-3.8232e-02	8.7896e+00		1.5200e+00	-1.2979e-01	1.8842e+01
8.4000e-01	-2.9696e-02	9.0224e+00		1.5400e+00	-1.3414e-01	1.9186e+01
8.6000e-01	-2.2344e-02	9.2584e+00		1.5600e+00	-1.3798e-01	1.9534e+01
8.8000e-01	-1.6128e-02	9.4976e+00		1.5800e+00	-1.4129e-01	1.9886e+01
9.0000e-01	-1.1000e-02	9.7400e+00		1.6000e+00	-1.4400e-01	2.0240e+01
9.2000e-01	-6.9120e-03	9.9856e+00		1.6200e+00	-1.4607e-01	2.0598e+01
9.4000e-01	-3.8160e-03	1.0234e+01		1.6400e+00	-1.4746e-01	2.0958e+01
9.6000e-01	-1.6640e-03	1.0486e+01		1.6600e+00	-1.4810e-01	2.1322e+01
9.8000e-01	-4.0800e-04	1.0742e+01		1.6800e+00	-1.4797e-01	2.1690e+01
1.0000e+00	0	1.1000e+01		1.7000e+00	-1.4700e-01	2.2060e+01
1.0200e+00	-3.9200e-04	1.1262e+01		1.7200e+00	-1.4515e-01	2.2434e+01
1.0400e+00	-1.5360e-03	1.1526e+01		1.7400e+00	-1.4238e-01	2.2810e+01
1.0600e+00	-3.3840e-03	1.1794e+01		1.7600e+00	-1.3862e-01	2.3190e+01
1.0800e+00	-5.8880e-03	1.2066e+01		1.7800e+00	-1.3385e-01	2.3574e+01
1.1000e+00	-9.0000e-03	1.2340e+01		1.8000e+00	-1.2800e-01	2.3960e+01
1.1200e+00	-1.2672e-02	1.2618e+01		1.8200e+00	-1.2103e-01	2.4350e+01
1.1400e+00	-1.6856e-02	1.2898e+01		1.8400e+00	-1.1290e-01	2.4742e+01
1.1600e+00	-2.1504e-02	1.3182e+01		1.8600e+00	-1.0354e-01	2.5138e+01
1.1800e+00	-2.6568e-02	1.3470e+01		1.8800e+00	-9.2928e-02	2.5538e+01
1.2000e+00	-3.2000e-02	1.3760e+01		1.9000e+00	-8.1000e-02	2.5940e+01
1.2200e+00	-3.7752e-02	1.4054e+01		1.9200e+00	-6.7712e-02	2.6346e+01
1.2400e+00	-4.3776e-02	1.4350e+01		1.9400e+00	-5.3016e-02	2.6754e+01
1.2600e+00	-5.0024e-02	1.4650e+01		1.9600e+00	-3.6864e-02	2.7166e+01
1.2800e+00	-5.6448e-02	1.4954e+01		1.9800e+00	-1.9208e-02	2.7582e+01
1.3000e+00	-6.3000e-02	1.5260e+01		2.0000e+00	0	2.8000e+01
1.3200e+00	-6.9632e-02	1.5570e+01		2.0200e+00	2.0808e-02	2.8463e+01
1.3400e+00	-7.6296e-02	1.5882e+01		2.0400e+00	4.3264e-02	2.8933e+01
1.3600e+00	-8.2944e-02	1.6198e+01		2.0600e+00	6.7416e-02	2.9409e+01
1.3800e+00	-8.9528e-02	1.6518e+01		2.0800e+00	9.3312e-02	2.9892e+01
1.4000e+00	-9.6000e-02	1.6840e+01		2.1000e+00	1.2100e-01	3.0382e+01
1.4200e+00	-1.0231e-01	1.7166e+01		2.1200e+00	1.5053e-01	3.0879e+01
1.4400e+00	-1.0842e-01	1.7494e+01		2.1400e+00	1.8194e-01	3.1382e+01
1.4600e+00	-1.1426e-01	1.7826e+01		2.1600e+00	2.1530e-01	3.1893e+01
1.4800e+00	-1.1981e-01	1.8162e+01		2.1800e+00	2.5063e-01	3.2411e+01
1.5000e+00	-1.2500e-01	1.8500e+01		2.2000e+00	2.8800e-01	3.2936e+01

Figure 1: The available information about the polynomial p .

It is easy to verify that

$$g(0.02) \approx 4, \quad g(-0.02) \approx 4 \tag{4}$$

This evidence is consistent with the conjecture that p has a double root at $x = 1$.

Problem 2 Consider the problem of computing the function $f : (0, \infty) \rightarrow \mathbb{R}$ given by

$$f(x) = \frac{e^x - 1}{x}. \quad (5)$$

1. (5 points) Show that $f(x) \rightarrow 1$ for $x \rightarrow 0_+$.

Solution The nature of f suggests that l'Hospital's rule should be applied. To that end, let $T(x) = e^x - 1$ and $N(x) = x$ for $x \in \mathbb{R}$. It is clear that T and N are differentiable and $T'(x) = e^x \rightarrow 1$ and $N'(x) = 1 \rightarrow 1$ as $x \rightarrow 0_+$. It follows, that $T'(x)/N'(x) \rightarrow 1$ as $x \rightarrow 0_+$. By l'Hospital's rule we can now conclude, that

$$f(x) = \frac{T(x)}{N(x)} \rightarrow 1, \quad x \rightarrow 0, \quad x > 0 \quad (6)$$

2. (5 points) Show that f is strictly increasing for $x > 0$.

Solution The function f is differentiable, because it is the quotient of two differentiable functions. Moreover, we have

$$f'(x) = \frac{e^x x - (e^x - 1)}{x^2} = \frac{e^x(x - 1) + 1}{x^2} > \frac{(x + 1)(x - 1) + 1}{x^2} = 1, \quad (7)$$

for $x > 0$. It follows, that f is strictly increasing for $x > 0$.

3. The following MATLAB commands have been used to generate Figure 2.

```
f=@(x)(exp(x)-1)./x;
x=linspace(1,2,1025)*2^(-48);
plot(x,f(x))
```

(5 points) Explain why it is immediately clear that this naive approach is unsuitable for practical computations!

Solution The plot is in obvious disagreement with the fact that f is strictly increasing for $x > 0$.

4. (5 points) Let u denote the unit round off. Give a formula which can be used to compute f without fear of catastrophic cancellation for all $x > 0$.

Solution From the *known* Taylor expansion of $x \rightarrow e^x$, we deduce that

$$f(x) = \sum_{n=1}^{\infty} \frac{x^{n-1}}{n!} = 1 + \frac{1}{2}x + \frac{1}{6}x^2 + \dots \quad (8)$$

Truncating this series after n terms, yields a family of approximations $\{p_m\}_{m=0}^{\infty}$ where

$$p_m(x) = \sum_{j=0}^m \frac{x^j}{(j+1)!} \quad (9)$$

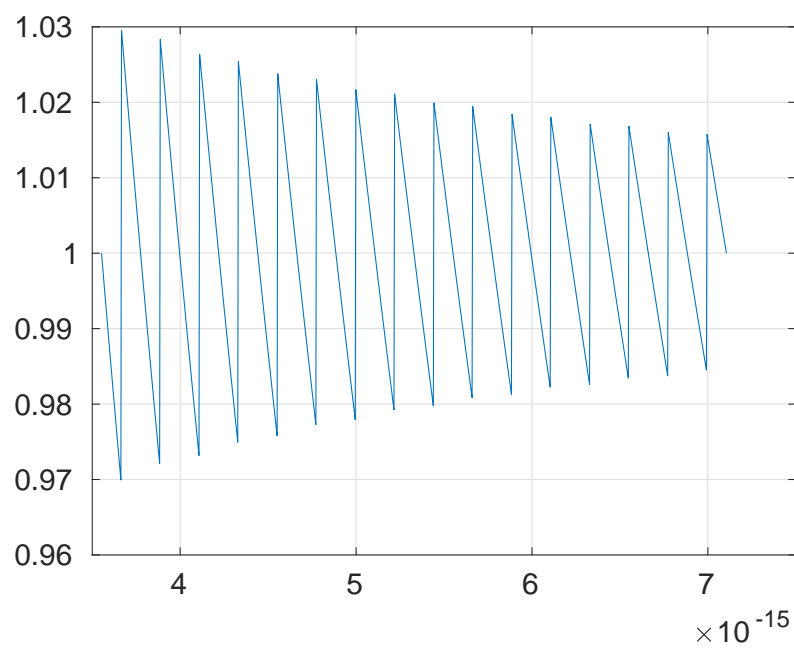


Figure 2: The graph of f as generated by the naive application of MATLAB commands.

There can be no subtractive cancellation for $x > 0$, because the polynomials have positive coefficients.

5. (5 points) Estimate the relative error when your formula is evaluated using floating point arithmetic and $0 < x < \sqrt{u}$.

Solution For such tiny values of x , we suspect that the higher order terms are negligible, and

$$f(x) \approx 1 + \frac{1}{2}x \quad (10)$$

Let $0 < x < \sqrt{u}$. By Taylor's formula, we have

$$f(x) - p_1(x) = \frac{f''(\xi)}{2}x^2, \quad 0 < \xi < x \quad (11)$$

We estimate $f''(\xi) \approx f''(0) = \frac{1}{3}$. It follows that

$$|f(x) - p_1(x)| \lesssim \frac{1}{3}u. \quad (12)$$

Since $f(x) > 1$, we have

$$\left| \frac{f(x) - p_1(x)}{f(x)} \right| \leq \frac{1}{3}u \quad (13)$$

Evaluating $y = p_1(x)$ can be done using Horner's method with a relative error which is less than $2u$, i.e.

$$\left| \frac{p_1(x) - \hat{y}}{p_1(x)} \right| \leq 2u \quad (14)$$

It follows that

$$\left| \frac{f(x) - \hat{y}}{f(x)} \right| \leq \left| \frac{f(x) - p_1(x)}{f(x)} \right| + \left| \frac{p_1(x) - \hat{y}}{f(x)} \right| \lesssim 2uu + \frac{1}{3}u < 3u. \quad (15)$$

Problem 3 Figure 3 and Figure 4 contain the available information about a piece of artillery. The range and the flight time of a shell is given as function of the elevation. The tables have been computed using the same method, but different time steps.

Elevation (degrees)	Range (meters)	Flight time (seconds)
0.00	0.000000e+00	0.000000e+00
5.00	5.715764e+03	1.075393e+01
10.00	9.314048e+03	2.017887e+01
15.00	1.189946e+04	2.871073e+01
20.00	1.387513e+04	3.660030e+01
25.00	1.541560e+04	4.399712e+01
30.00	1.659769e+04	5.098788e+01
35.00	1.744766e+04	5.761594e+01
40.00	1.796164e+04	6.389240e+01
45.00	1.811640e+04	6.980335e+01
50.00	1.787705e+04	7.531546e+01
55.00	1.720448e+04	8.038100e+01
60.00	1.606361e+04	8.494279e+01
65.00	1.443249e+04	8.893927e+01
70.00	1.231103e+04	9.230930e+01
75.00	9.728114e+03	9.499660e+01
80.00	6.746499e+03	9.695352e+01
85.00	3.460122e+03	9.814325e+01
90.00	2.451075e-12	9.854249e+01

Figure 3: Range and flight time computed using time step $dt = 0.1$ and `method = 'rk2'`.

- (5 points) What evidence do you find to support the conjecture that these tables relates to shells fired into an atmosphere without any wind?

Solution The range function appear to be unimodal with a global maximum in the vicinity of $\theta = 45^\circ$. However the range function is *not* symmetrical around the $\theta = 45^\circ$. This proves that the shells are not fired in a vacuum. Since the elevation $\theta = 90^\circ$ results in a range which is essentially zero, this strongly suggests that there is no wind.

- (5 points) Estimate the range of the *real* gun as accurately as the data will allow when the elevation is 45 degrees.

Solution There are only two relevant values, namely $r_h = 1.811640 \times 10^4$ and $r_{2h} = 1.811633 \times 10^4$ corresponding to the timestep $h = 0.1$. Richardson's error estimate is

$$E_h = \frac{r_h - r_{2h}}{3} \approx 2.3333 \times 10^{-2}, \quad (16)$$

Elevation (degrees)	Range (meters)	Flight time (seconds)
0.00	0.000000e+00	0.000000e+00
5.00	5.715535e+03	1.075354e+01
10.00	9.313877e+03	2.017852e+01
15.00	1.189932e+04	2.871041e+01
20.00	1.387503e+04	3.660001e+01
25.00	1.541551e+04	4.399684e+01
30.00	1.659761e+04	5.098761e+01
35.00	1.744758e+04	5.761567e+01
40.00	1.796157e+04	6.389212e+01
45.00	1.811633e+04	6.980306e+01
50.00	1.787698e+04	7.531515e+01
55.00	1.720441e+04	8.038068e+01
60.00	1.606355e+04	8.494246e+01
65.00	1.443243e+04	8.893892e+01
70.00	1.231097e+04	9.230893e+01
75.00	9.728070e+03	9.499622e+01
80.00	6.746468e+03	9.695313e+01
85.00	3.460106e+03	9.814285e+01
90.00	2.451063e-12	9.854209e+01

Figure 4: Range and flight time computed using time step $dt = 0.2$ and method `'rk2'`.

i.e., less than an inch. The “best” we can do is to return the value $r_h + E_h$, but our position is rather weak. We have assumed that the method `'rk2'` delivers approximations which are second order accurate, i.e. $p = 2$. We have assumed that the error estimate E_h is reliable.

3. (5 points) What auxiliary information do you require before you can vouch for the validity of your error estimate?

Solution We can not vouch for the reliability of Richardson’s estimate without observing the behavior of Richardson’s fraction $F_h = \frac{r_{2h} - r_{4h}}{r_h - r_{2h}}$ for multiple values of h . To that end, we require approximation $r_{2^j h}$ computed for a suitably large range of values of j , say, $j \in \{0, 1, 2, \dots, 5\}$.

4. (10 points) A skilled crew can reload in less than 20 seconds. Explain why it is theoretically possible for a skilled crew to fire two different shells, such that they detonate *simultaneously* on the *same* target located 16000 meters from the gun.

Solution The target is within range of the gun, hence there is both a low (fast) and a high (slow) trajectory to the target. Assuming the flight times can be trusted, then it is clear that the flight time along the low

trajectory is (certainly) less than 51 seconds, and the flight time along the high trajectory is (certainly) greater than 84 seconds. The difference is greater than 33 seconds. The trained crew will fire a shell along the high trajectory, reload, change elevation and fire the second shell along the low trajectory at the appropriate time, i.e. at least 33 seconds after the first shell.

Component	time instant(s) with questionable error estimate
1	none
2	30
3	none
4	none

Problem 4 Figure 5 and 6 have been generated by applying Richardson's techniques to the problem of computing the trajectory $\gamma(t) = (x(t), y(t), x'(t), y'(t))$ of a specific artillery shell. The value of the smallest time step h used to integrate the trajectories has been lost.

1. (5 points) Identify all error estimates of questionable validity.

Solution It is a matter of examining Richardson's fraction. We can quickly see that almost all fractions are close to 2, which strongly suggests that the trajectory has been computed using a first order method, i.e., $p = 1$. The fundamental problem is to *quickly* identify the error estimates which are not reliable. Examining the last two values of Richardson's fraction is the fastest way to go. It is a matter of computing $\frac{2-F_{2h}}{2-F_h}$ and comparing this value to 2. There is only *one* sequence which does not follow the standard pattern, i.e., (eventual) monotone convergence towards a power of 2 at a fixed rate. Specifically, the data related to $y(30)$ (Figure 5, Component 2, $t = 30.00$). There is not enough data to determine if the error estimate is reliable. The standard cure is to compute more approximations using smaller stepsizes and form the corresponding fractions.

2. Let $T = y'(30)$ and let A_h denote our approximation.
 - (a) (10 points) What evidence do you find to support the conjecture that the error $T - A_h$ obeys an asymptotic error expansion of the form

$$T - A_h = \alpha h + \beta h^2 + O(h^r), \quad 2 < r.$$

Remark 1 We are limiting our attention to Figure 6, Component 4, and $t = 30.00$.

Solution Consider the more general hypothesis

$$T - A_h = \alpha h^p + \beta h^q + O(h^r), \quad 0 < p < q < r. \quad (17)$$

If this hypothesis is true, then we must have $F_h \rightarrow 2^p$ and monotonically so, when h is small enough. Moreover, $F_h - 2^p = O(h^{q-p})$. Along the specific row Richardson's fractions are seen to decay monotonically towards 2. This suggests that the order of the dominant error term is $p = 1$. Moreover, it is clear that $F_h - 2 = O(h)$ which implies $q - p = 1$. Equivalently, $q = 2$.

(b) (5 points) Determine the sign of α .

Solution Richardson's error estimate is $E_h = \frac{A_h - A_{2h}}{2^p - 1} \approx \alpha h^2$. From the specific row of the table, we see that $E_h > 0$. This implies $\alpha > 0$.

(c) (5 points) Determine the sign of β .

Solution Since Richardson's fraction F_h are decaying towards 2 we must have $\frac{\beta}{\alpha} > 0$. Since $\alpha > 0$ we must have $\beta > 0$.

Data related to component 1

time	Ah : approximation	F_(16h)	F_(8h)	F_(4h)	F_(2h)	F_(1h)	Error estimate
0.00	0.000000000000000e+00	NaN	NaN	NaN	NaN	NaN	0.000000000000000e+00
5.00	2.4867256112632022e+03	2.167413	2.080868	2.038311	2.018969	2.009393	-1.2811341210631326e+00
10.00	4.1563802797672852e+03	2.099443	2.055001	2.024628	2.012419	2.006136	-1.1096434567466531e+00
15.00	5.4665160370997037e+03	2.003606	2.020213	2.004938	2.003054	2.001446	-6.1627109365781507e-01
20.00	6.6509365167841861e+03	1.487855	1.875961	1.917218	1.963279	1.981602	-1.3336331789378164e-01
25.00	7.7609608314982879e+03	2.313964	2.115010	2.068346	2.032687	2.016504	3.1758884329701687e-01
30.00	8.8115820962937523e+03	2.199119	2.077679	2.043420	2.020883	2.010485	7.3581948014725640e-01
35.00	9.8098087323611326e+03	2.167571	2.067578	2.036813	2.017775	2.008906	1.1215242708094593e+00
40.00	1.0757610953208115e+04	2.153678	2.063227	2.033952	2.016434	2.008226	1.4754572148831357e+00
45.00	1.1653374785789472e+04	2.146807	2.061210	2.032571	2.015794	2.007900	1.7988011010584160e+00
50.00	1.2492934932197231e+04	2.143552	2.060419	2.031953	2.015514	2.007757	2.0932932500381867e+00

Data related to component 2

time	Ah : approximation	F_(16h)	F_(8h)	F_(4h)	F_(2h)	F_(1h)	Error estimate
0.00	0.000000000000000e+00	NaN	NaN	NaN	NaN	NaN	0.000000000000000e+00
5.00	2.4012333248677692e+03	2.150053	2.072028	2.034187	2.016907	2.008372	-1.6038307168992105e+00
10.00	3.7955517997170259e+03	2.083108	2.044642	2.020280	2.010187	2.005037	-1.6109493672715871e+00
15.00	4.6598034076994236e+03	2.016296	2.017316	2.006037	2.003307	2.001611	-1.2421708048068467e+00
20.00	5.2019492427161349e+03	1.929872	1.984337	1.988432	1.994905	1.997426	-8.3807301043907501e-01
25.00	5.4702136646835488e+03	1.668836	1.892387	1.940550	1.972456	1.986325	-4.2950104114788701e-01
30.00	5.4836939336948790e+03	5.728888	6.698715	-7.409434	0.617207	1.487261	-1.8210754560641362e-02
35.00	5.2559164107667302e+03	2.518487	2.226088	2.127074	2.062789	2.031770	3.9745405791927624e-01
40.00	4.7992752468479612e+03	2.335120	2.144044	2.078088	2.038256	2.019225	8.2016818876218167e-01
45.00	4.1278893750924353e+03	2.266111	2.114541	2.061074	2.029865	2.014970	1.2521711129384130e+00
50.00	3.2596858802413149e+03	2.227816	2.098392	2.051925	2.025377	2.012702	1.6940507297545082e+00

Figure 5: Data related to the position of the shell, i.e. $(x(t), y(t))$.

Data related to component 3

time	Ah : approximation	F_(16h)	F_(8h)	F_(4h)	F_(2h)	F_(1h)	Error estimate
0.00	5.9396969619669994e+02	NaN	NaN	NaN	NaN	NaN	0.0000000000000000e+00
5.00	3.7551591305403076e+02	2.156452	2.068214	2.034672	2.016950	2.008438	2.1726174222237660e-01
10.00	2.7934473545198972e+02	2.097882	2.044026	2.022575	2.011090	2.005534	2.1576334520966611e-01
15.00	2.4350505653754701e+02	2.064791	2.028469	2.014156	2.006897	2.003429	1.4190410779713147e-01
20.00	2.2696935056525186e+02	2.068922	2.031021	2.015520	2.007588	2.003776	1.2289333044878958e-01
25.00	2.1427691055286675e+02	2.068713	2.030709	2.015371	2.007505	2.003733	1.1073697609248256e-01
30.00	2.0345524847359059e+02	2.069712	2.030984	2.015530	2.007576	2.003769	1.0150842069580790e-01
35.00	1.9335080491550150e+02	2.071987	2.031837	2.015983	2.007791	2.003876	9.4303887096401695e-02
40.00	1.8315647040746524e+02	2.075245	2.033146	2.016670	2.008122	2.004041	8.8680760929236158e-02
45.00	1.7229227649036287e+02	2.078976	2.034676	2.017476	2.008512	2.004235	8.4320257035216173e-02
50.00	1.6035834541225086e+02	2.082721	2.036227	2.018297	2.008911	2.004434	8.1031723701016745e-02

Data related to component 4

time	Ah : approximation	F_(16h)	F_(8h)	F_(4h)	F_(2h)	F_(1h)	Error estimate
0.00	5.9396969619669983e+02	NaN	NaN	NaN	NaN	NaN	0.0000000000000000e+00
5.00	3.3330830724420468e+02	2.151551	2.066624	2.033713	2.016495	2.008210	2.4121319808892849e-01
10.00	2.0347653733758088e+02	2.096675	2.043708	2.022277	2.010950	2.005462	2.4940275054626682e-01
15.00	1.3014392088345306e+02	2.069727	2.031019	2.015577	2.007623	2.003795	1.9963337165842177e-01
20.00	7.3473584343104633e+01	2.069457	2.031101	2.015624	2.007650	2.003808	1.8795486079237378e-01
25.00	2.1340253105106484e+01	2.068102	2.030418	2.015263	2.007469	2.003717	1.8131350511454514e-01
30.00	-2.7843684493860440e+01	2.067098	2.029952	2.015016	2.007346	2.003655	1.7674689540964295e-01
35.00	-7.4566502453673621e+01	2.065773	2.029387	2.014721	2.007203	2.003583	1.7305889376771688e-01
40.00	-1.1866866328488541e+02	2.063511	2.028439	2.014239	2.006969	2.003467	1.6922344123553046e-01
45.00	-1.5952561265887957e+02	2.059801	2.026878	2.013452	2.006588	2.003278	1.6425284203688761e-01
50.00	-1.9617052089851981e+02	2.054021	2.024428	2.012226	2.005994	2.002984	1.5712509505982553e-01

Figure 6: Data related to the velocity of the shell, i.e. $(x'(t), y'(t))$.