Remark 1 This one possible way of solving the exam and receiving a maximal score. I have written more text than I expect students to do. This is due to the fact that the old exams are used to train new students. It is very likely that there a couple of typographical errors in the text. Do not hesitate to contact me to get confirmation.

Problem 1 Let I be an interval and let $h : I \to \mathbb{R}$ be a differentiable function, such that the equation

$$h(x) = 0 \tag{1}$$

has exactly one solution $r \neq 0$. Moreover, $h'(r) \neq 0$.

1. (5 points) Give a real life example of a function h where a small residual is much more important than a small error.

   Solution Consider 2 dimensional projectile motion. Given the range r to the target the problem consists of computing the elevation θ which will land the round on the target, i.e. $r = x(\theta, \tau(\theta))$. Let $\hat{\theta}$ be the computed angle. The error is $\theta - \hat{\theta}$, but in real life we primarily care about the size of the residual, i.e. $R = r - x(\hat{\theta}, \tau(\hat{\theta}))$. If $|R|$ is less than the kill radius, then the target is destroyed and we can move on to the next problem.

2. (10 points) Explain, why it is theoretically impossible to compute $h(x)$ with a small relative error using floating point arithmetic, when we are very close to the true root r.

   Solution The condition number of h limits how accurate we can compute h. Specifically, if $x \neq r$, then there is a ξ between x and $x + \Delta x$ such that

   $$\frac{h(x + \Delta x) - h(x)}{h(x)} = \frac{h'(\xi)x}{h(x)} \frac{\Delta x}{x} \tag{2}$$

   It follows that

   $$\left| \frac{h(x + \Delta x) - h(x)}{h(x)} \right| \approx \kappa_h(x) \left| \frac{\Delta x}{x} \right|, \quad \kappa_h(x) = \left| \frac{\Delta x}{x} \right| \tag{3}$$

   In our situation we have

   $$\kappa_h(x) \to \infty, \quad x \to r, \quad x \neq r \tag{4}$$

   for the very simple reason that $r \cdot h'(r) \neq 0$.

3. (10 points) Explain, why it is possible for a real implementation of the bisection algorithm to return an interval $[a_n, b_n]$ which does not contain r, even though the initial interval $[a_0, b_0]$ contains r.

   Solution The bisection algorithm hinges on our ability to correctly compute the sign of h at select points x_j. If we get too close to the root r, then by the previous question, we can not be certain that the $h(x_j)$ is computed with a relative error which is bounded by 1, therefore we can not be certain that the sign is computed correctly. In popular terms, if a computer is not forced to give correct answers, then it will exploit this freedom and generate garbage.

Problem 2 Consider the problem of two dimensional projectile motion. Figure 1 contains a partial artillery table for a particular gun fired under standard conditions, i.e. constant gravity, constant temperature, homogene atmosphere, and no wind.

Elevation theta (degrees)	Range r (meters)	Flight time tau (seconds)	Derivative dr/dtheta (meters/radian)
4.00	6286	10.15	66754
8.00	10053	18.92	43813
12.00	12632	26.75	31091
16.00	14492	33.88	22668
20.00	15847	40.46	16416
24.00	16812	46.59	11395
28.00	17455	52.32	7128
32.00	17819	57.71	3349
36.00	17931	62.77	-99
40.00	17810	67.54	-3317
44.00	17471	72.01	-6372
48.00	16923	76.20	-9310
52.00	16173	80.09	-12162
56.00	15226	83.68	-14951
60.00	14087	86.95	-17691
64.00	12757	89.88	-20385
68.00	11242	92.46	-23026
72.00	9544	94.66	-25591
76.00	7671	96.46	-28031
80.00	5635	97.83	-30254
84.00	3456	98.76	-32104
88.00	1168	99.23	-33313

Figure 1: A partial ballistics table for a gun fired under standard conditions.

1. (5 points) Explain, why you can be virtually certain that the maximal range of the gun corresponds to an elevation between 32 and 36 degrees.

   Solution Let g denote the function which converts from degrees to radian. Let $r(\theta)$ denote the range as a function of the elevation θ. The second column of the table suggests that the range function

   $$r(\theta) \leq 17900 \tag{5}$$

   for $\theta < g(32)$ or $g(40) < \theta$. Assuming that r is continuous there must be a global maximum θ_0 in the interval $[32, 40]$ and we must have $r'(\theta_0) = 0$. The fourth column shows that r' has a zero between 32 and 36 degrees and it also suggests that r' is strictly monotone decreasing. It follows, that θ_0 is between $g(32)$ and $g(36)$. We stress that our argument hinges on the continuity of r'.

Remark 2 It is possible to show that r is smooth, but this is entirely beyond the scope of this class.

2. (5 points) Estimate one of two possible flight times to a target $r = 8000$ meters. Your relative error must less than 2 procent.

   Solution By examining the table we that there is a firing solution $\theta \in (g(76), g(80)$ and that the corresponding flight time $\tau \in (a, b) = (96.46, 97.83)$. Without additional information our best estimate for the flight time is the average $\hat{\tau} = \frac{a+b}{2} = 97.145$. The error is bounded by $e = \frac{b-a}{2} = 0.685$ and the relative error is bounded by $\frac{e}{a} \approx 7 \times 10^{-3}$ which is much less than the required tolerance $\tau = 0.02$.

3. (10 points) Use Newton's method to compute the elevation corresponding to both the high and the low trajectory to a target at $r = 10000$ meters.

   Hint: Do not forget to convert between degrees and radians as needed!

   Solution This problem is diabolical, because the lack of information limit what we can do. However, it is critical to keep in mind what the real goal is! Here we are not interested in computing the elevation with a relative error which is comparable to the unit round-off error. Instead we to put a round within the kill radius of the target! Let g denote the function which converts from degrees to radian.

   The low trajectory It is clear that elevation corresponding to the low trajectory is extremely close to 8 degrees, in fact with $\theta_0 = g(8)$, then

$$\begin{aligned}\theta_1 = \theta_0 - \frac{r(\theta_0) - r}{r'(\theta_0)} &= \theta_0 - \frac{53}{43813} \\ &= 0.13962634 - 0.0012096866 = 0.13841665 \approx 7.93 \text{ degrees.}\end{aligned}$$

   Now the error of Newton's method converges quadratically, if we start sufficiently close to a root. Similarly, the residual also converges quadratically. Here the initial residual is about 53 m and so we expect the new residual to be around 7 m, much less than the standard kill radius of 30 m which we have used in the class.

   The high trajectory It is clear that the elevation corresponding to the high trajectory to the target is between 68 and 72 degrees and closer to 72 degrees. With $\theta_0 = g(72)$ we compute

$$\begin{aligned}\theta_1 = \theta_0 - \frac{r(\theta_0) - r}{r'(\theta_0)} &= 1.2566371 - \frac{-456}{-25591} \\ &= 1.2566371 - 0.017818764 = 1.2388183 \approx 71.0 \text{ degrees.}\end{aligned}$$

   The initial residual is 456 meters and so we hope/expect the new residual to be around 21 meters.

   In both cases it is likely that the target will be destroyed in the first attempt and no further iterations are necessary. In partice, one would fire two small sets of shells towards the target using angles grouped around 7.93 and 71.0 degrees, just to be sure.

It is possible to execute an approximation of Newton's method in this context using interpolation to approximate the missing values of the derivative. However, this problem specifically requested Newton's method.

4. (5 points) Explain why it is theoretically possible to have two rounds hit the same target at $r = 10000$ meters *simultanously* using a *single* gun provided the crew can reload in less than 60 seconds.

   Solution By manual inspection of the table we see that the flight time to the target using the low trajectory is less than 18.92 seconds and the flight time to the target using the high trajectory is greater than 92.46 seconds. The difference between these two flight times is greater than 73.46 seconds. Since the crew can reload in 60 seconds it is at least theoretically possible to first fire a round along the high trajectory, change the elevation to the low trajectory, reload, and finally fire the second round, so that the two rounds impact the target simultanously.

   Remark 3 By varying the muzzle velocity it is possible to have the same gun fire several rounds which land on the same target simultanously. This technique is known as "Multiple Rounds Simultanous Impact" and abbreviated MRSI.

Problem 3 Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} \frac{x-\cos(x)\sin(x)}{x^3}, & x \neq 0 \\ \frac{2}{3}, & x = 0. \end{cases} \tag{6}$$

1. (5 point) Show that f is differentiable for $x \neq 0$.

   Solution The function f is built from continuous functions such as $x \to x^n$, $x \to \cos(x)$, $x \to \sin(x)$ using the basic arithmetic operations. It follows that f is automatically continuous everywhere it is defined. The only potential problem is $x = 0$, which is not under consideration in this question.

2. (5 point) Show that $f(x) \to f(0)$ for $x \to 0$, $x \neq 0$.

   Solution We have

   $$T(x) = x - \cos(x)\sin(x) \to 0, \quad x \to 0, \quad x \neq 0, \tag{7}$$
   $$N(x) = x^3 \to 0, \quad x \to 0, \quad x \neq 0. \tag{8}$$

   We also have

   $$T'(x) = 1 + \sin(x)^2 - \cos(x)^2 \to 0, \quad x \to 0, \quad x \neq 0, \tag{9}$$
   $$N'(x) = 3x^2 \to 0, \quad x \to 0, \quad x \neq 0, \quad x \neq 0. \tag{10}$$

   We also have

   $$T^{(2)}(x) = 4\sin(x)\cos(x) \to 0, \quad x \to 0, \quad x \neq 0, \tag{11}$$
   $$N^{(2)}(x) = 6x \to 0, \quad x \to 0, \quad x \neq 0, \quad x \neq 0. \tag{12}$$

   We also have

   $$T^{(3)}(x) = 4(\cos(x)^2 - \sin(x)^2) \to 4, \quad x \to 0, \quad x \neq 0, \tag{13}$$
   $$N^{(3)}(x) = 6 \to 6, \quad x \to 0, \quad x \neq 0, \quad x \neq 0. \tag{14}$$

   Applying l'Hospital's rule 3 times we find that

   $$f(x) = \frac{T(x)}{N(x)} \to \frac{4}{6} = \frac{2}{3} = f(0). \tag{15}$$

   It follows that f is continuous also at $x = 0$.

3. (5 point) The following `MATLAB` commands

   ```
   >> f=@(x)(x-cos(x).*sin(x))./x.^3;
   >> x=linspace(-1,1,1025)*2^-23;
   >> plot(x,f(x));
   ```

   have been used to generate the graph given in Figure 2. Explain, why you can immediately conclude that this graph is not an accurate representation of the mathematical reality and we cannot rely on the above implementation of f.

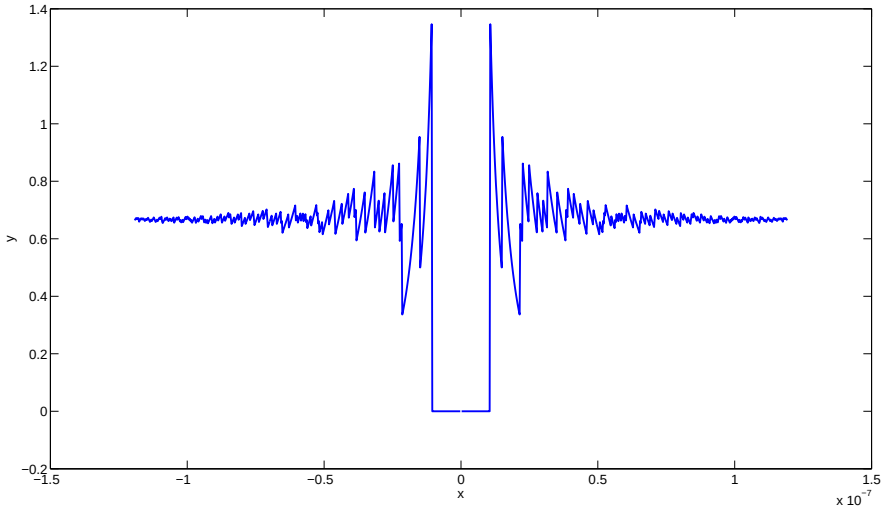

Figure 2: A plot of the function $x \to f(x)$ using a naive `MATLAB` implementation.

Solution We have just show that that f is continuous at $x = 0$ with $f(0) = \frac{2}{3}$. The naive plot impliest that $f(x) = 0$ in a neighborhood around $x = 0$. Therefore the plot is obviously wrong. Moreover, the erratic oscilations are not typical of a differentiable function such as f. They are, however, characteristic of the effects of subtractive cancellation.

4. (5 point) Show that f can be computed using the *alternating* series

$$f(x) = 4 \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+3)!} (2x)^{2j}. \tag{16}$$

Hint: You are free to exploit the identity

$$2 \sin(x) \cos(x) = \sin(2x), \tag{17}$$

as well as the Taylor series expansion for $x \to \sin(x)$ at $x = 0$, i.e.

$$\sin(x) = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} x^{2j+1}. \tag{18}$$

Solution Following the hint we have

$$\begin{aligned} \sin(x)\cos(x) = \frac{1}{2}\sin(2x) &= \frac{1}{2} \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} (2x)^{2j+1} \\ &= x + \frac{1}{2} \sum_{j=1}^{\infty} \frac{(-1)^j}{(2j+1)!} (2x)^{2j+1} = x - \frac{1}{2} \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+3)!} (2x)^{2j+3} \\ &= x - 4 \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+3)!} (2x)^{2j} x^3 \end{aligned} \tag{19}$$

It follows immediately that

$$f(x) = 4\sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+3)!}(2x)^{2j}. \tag{20}$$

5. (5 point) Let $T_n(x)$ denote the polynomial

$$T_n(x) = 4\sum_{j=0}^{n} \frac{(-1)^j}{(2j+3)!}(2x)^{2j}, \tag{21}$$

and consider $T_n(x)$ as an approximation of $f(x)$. Show that

$$\frac{|f(x) - T_0(x)|}{|f(x)|} \leq 10^{-14}, \tag{22}$$

for all $x \in [-2^{-23}, 2^{-23}]$.

Hint: Alternating series converge when the absolute value of the terms are decreasing monotonically. The truncation error is closely related to the absolute value of the first term which is neglected.

Solution Since $|2x| \leq 1$ the absolute value of the terms are decreasing monotonically. The size of the error is less than the first term which is neglected. It follows, that the error can be bounded as follows

$$|f(x) - T_0(x)| \leq \frac{4}{7!}(2x)^2 \leq \frac{4}{7!}(2 \times 2^{-23})^2 < 4.52 \times 10^{-17} < 6 \times 10^{-17} \tag{23}$$

As for the relative error we it is easy to see that we are meeting our tolerance. From

$$|T_0(x)| - |T_0(x) - f(x)| \leq ||T_0(x)| - |T_0(x) - f(x)|| \leq |f(x)| \tag{24}$$

it follows that

$$\begin{aligned} \frac{|f(x) - T_0(x)|}{|f(x)|} &\leq \frac{|f(x) - T_0(x)|}{|T_0(x)| - |T_0(x) - f(x)|} \\ &\lesssim \frac{6 \times 10^{-17}}{\frac{2}{3}} = 9 \times 10^{-17} \ll 10^{-14}. \end{aligned} \tag{25}$$

Problem 4 An unknown function $g : \mathbb{R} \to \mathbb{R}$ has been differentiated numerically using an unknown numerical method $D = D(h)$ at the point $x = 2$ and step sizes

$$h = h(k) = h_0 \times 2^{-k}, \quad k = 0, 1, 2, \ldots 19, \quad h_0 = \frac{3}{2}. \tag{26}$$

It is known that g is infinitely often differentiable at every point $x \in \mathbb{R}$. Figure 3 contains the computed approximations of $g'(2)$ as well as some auxiliary numbers.

k	Dh	Fractions (D2h-D4h)/(Dh-D2h)	Error estimate (Dh-D2h)/(2^p-1)
0	-1.8671501966806722e-01	0.0000000000000000e+00	0.000000e+00
1	-2.2506471579681531e-01	0.0000000000000000e+00	-1.278323e-02
2	-2.3563099682161945e-01	3.6294412422613886e+00	-3.522094e-03
3	-2.3833698477779799e-01	3.9047775510893512e+00	-9.019960e-04
4	-2.3901755946574385e-01	3.9760337904218725e+00	-2.268582e-04
5	-2.3918795880368884e-01	3.9939984283598209e+00	-5.679978e-05
6	-2.3923057462996411e-01	3.9984989812080878e+00	-1.420528e-05
7	-2.3924122958621533e-01	3.9996247070843687e+00	-3.551652e-06
8	-2.3924389338776564e-01	3.9999061679227328e+00	-8.879339e-07
9	-2.3924455934204994e-01	3.9999765946503314e+00	-2.219848e-07
10	-2.3924472583085313e-01	3.9999944233381335e+00	-5.549627e-08
11	-2.3924476745310130e-01	3.9999954467843857e+00	-1.387408e-08
12	-2.3924477785870599e-01	3.9999836090244298e+00	-3.468535e-09
13	-2.3924478046001241e-01	4.0001456969398443e+00	-8.671021e-10
14	-2.3924478111014955e-01	4.0011657017377171e+00	-2.167124e-10
15	-2.3924478127264592e-01	4.0009331205529239e+00	-5.416546e-11
16	-2.3924478131448268e-01	3.8840577498726221e+00	-1.394559e-11
17	-2.3924478132539662e-01	3.8333333333333335e+00	-3.637979e-12
18	-2.3924478132297131e-01	-4.5000171662031789e+00	8.084367e-13
19	-2.3924478132782193e-01	-5.0000000000000000e-01	-1.616873e-12

Figure 3: The results obtained by differentiating the function $x \to g(x)$ numerically using the method $D = D(h)$ at the point $x = 2$.

1. (5 points) Explain, why it is likely the numerical method D has order $p = 2$.

   Solution We know next to nothing about the numerical method, but it appears to converge as $h \to 0$. Therefore it is not entirely impossible that it supports an asymptotic error expansion of the form

   $$g'(x) - D_h(x) = \alpha(x)h^p + \beta(x)h^q + O(h^r) \tag{27}$$

   where $p < q < r$ are positive integers. If such an expansion exists, then the standard analysis shows that

   $$\frac{D_{2h} - D_{4h}}{D_h - D_{2h}} \to 2^p, \quad h \to 0, \quad h \neq 0 \tag{28}$$

   Moreover, the convergence will be monotonic. This is exactly the behavior we observe for $k = 2, 3, \ldots, 11$, where Richardson's fractions approach the value 4. This strongly indicates that the order is $p = 2$.

2. (5 points) Explain, why Richardson's fractions deviate from 4 for small values of k.

   Solution The deviation is cause by the presence of higher order terms in the asymptotic error expansion. Their effect can not be ignored for large values of $h = h_0 2^{-k}$, i.e. small values of k.

3. (5 points) Explain, why the computed value of Richardson's fraction deviates from 4 for large values of k.

   Solution If k is large, then h is small and the three numbers D_h, D_{2h} and D_{4h} are close to the real number $g'(x)$. It follows that we will experience subtractive cancellation when computing Richardson's fraction. In popular terms we will divide garbage with garbage and we have no right to expect to get a value which is close to 4. The larger k gets, the worse the problem becomes, a fact which is also reflected in the table.

4. (5 points) Find the range of k where the computed value of Richardson's fraction behaves as if it had been computed in exact arithmetic.

   Solution For $k = 2, 3, \ldots, 11$ the computed fractions are increasing towards 4 from below. This is consistent with the theory for the exact fractions. The value at $k = 13$ is clearly wrong, because it has jumped to the other side, i.e. it is stricly greater than 4. The value at $k = 12$ is smaller than the value of $k = 11$ which violates the theory as well. Therefore the requested range is $k = 2, 3, \ldots, 11$.

5. (5 points) Explain, why the number of correct significant figures of the computed error estimate is in all likelihood maximal at $k = 11$.

   Solution This is a matter of experience. In the past we have seen the number of correct digits peak at the last index where the computed fractions behaved as predicted by the theory for the real fractions. There is nothing in the observed behavior of the computed values of Richardson's fraction to suggest any deviation from this behavior.