Problem 1 Consider the function $f : [1, \infty) \to \mathbb{R}$

$$f(x) = \sqrt{x^4 + 1} - \sqrt{x^4 - 1}$$

1. (5 points) Show that f is differentiable and strictly decreasing for all $x > 1$.

   Solution It is well known that polynomials and square roots are differentiable functions. Therefore f is built from differentiable functions using a finite number of arithmetic operations and function compositions. Therefore f is differentiable at every point where all the components are differentiable, i.e. for $x > 1$. Moreover we have

   $$f'(x) = 2x^3 \left(\frac{1}{\sqrt{x^4 + 1}} - \frac{1}{\sqrt{x^4 - 1}} \right) < 0$$

   from which it follows that f is strictly decreasing.

2. (5 points) The `MATLAB` commands

```
>>x=single(linspace(10,100,201));
>>f=sqrt(x.^4+1)-sqrt(x.^4-1);
>>plot(x,f)
```

   followed by a few purely cosmetic commands have generated the graph displayed in Figure 1. Which features of this graph have nothing to do with reality?

   Solution We know that the function is strictly decreasing and it is also clear that it is strictly positive. It follows that the oscillations which appear to start around $x = 24$ as well as the constant behavior for x greater than about 76 has nothing to do with reality.

3. (5 points) Why did the `MATLAB` commands fail to produce a reliable plot?

   Solution The expression for f suffers from catastrophic cancellation for "large" values of x. In fact, since $\mathrm{fl}(1+2^{24}) = 2^{24}$ the term $+1$ is irrelevant for $x > 2^6 = 64$. Moreover, while $\mathrm{fl}(2^{24} - 1) = 2^{24} - 1$ we certainly have $\mathrm{fl}(2^{25} - 1) = 2^{25}$, and so f is evaluated as zero for all $x > \sqrt[4]{2^{25}} > 76$.

4. (5 points) Why is catastrophic cancellation not an issue for the interval

   $$1 < x < \sqrt[4]{\frac{5}{3}}.$$

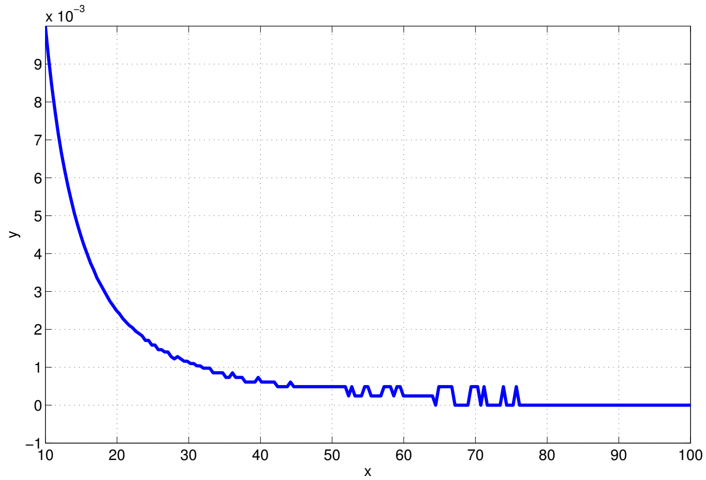

Figure 1: The naive application of `MATLAB` to the problem of computing f

Solution In general, a subtraction $a - b$ is entirely safe if $a > 2b > 0$. In our case we are dealing with subtractions of the form $d(x) = a(x) - b(x)$ where

$$a(x) = \sqrt{x^4 + 1}, \quad \text{and} \quad b(x) = \sqrt{x^4 - 1}$$

When is $a(x) > 2b(x)$? But we have $x > 1$ and so

$$\sqrt{x^4 + 1} > 2\sqrt{x^4 - 1} \Leftrightarrow x^4 + 1 > 4x^4 - 4 \Leftrightarrow 5 > 3x^4 \Leftrightarrow x < \sqrt[4]{\frac{5}{3}}.$$

5. (5 points) Find a numerically reliable way to evaluate $f(x)$ for all $x \geq 1$ using `MATLAB`.

Solution We have to find an expression which is mathematically equivalent to the definition of f, but which does not cancel catastrophically. We have

$$\sqrt{x^4 + 1} - \sqrt{x^4 - 1} = \frac{x^4 + 1 - (x^4 - 1)}{\sqrt{x^4 + 1} + \sqrt{x^4 - 1}} = \frac{2}{\sqrt{x^4 + 1} + \sqrt{x^4 - 1}}$$

and the last expression does not cancel catastrophically for large value of x. The term $x^4 - 1 \approx 0$ will cancel catastrophically for $x \approx 1$, but this error is irrelevant, because because the other term, i.e. $x^4 + 1 \approx 2$ is much larger.

Problem 2 Consider the function $g : \mathbb{R} \rightarrow \mathbb{R}$ given by

$$g(x) = x^3 - x^2 - 4x + 1.$$

A very crude plot of the graph of g can be found in Figure 2.

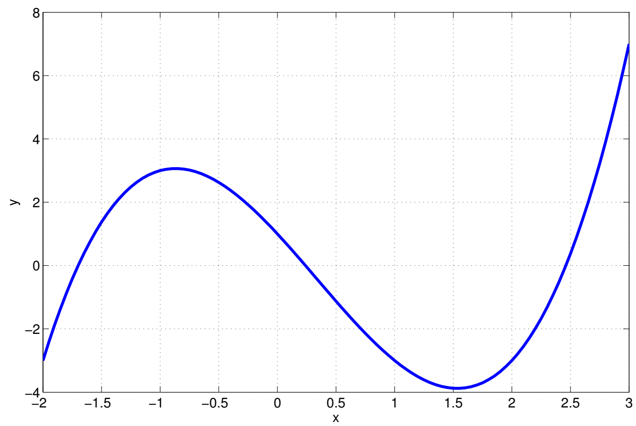

Figure 2: A crude plot of the graph of g.

1. (5 points) Explain why you can be absolutely certain that g has exactly three distinct zeros even though we are not certain that the graph can be trusted.

   Solution By a direct computation we find

   $$p(-2) = -3, p(-1) = 3, p(1) = -3, p(3) = 7.$$

   Since p is a polynomial it is a continuous function. Therefore there is at least one root in each of the three disjoint intervals

   $$(-2, -1), (-1, 1), (1, 3)$$

   Since p is a polynomial of degree 3 there can be no other roots than these three real numbers.

2. (10 points) Newton's method has been applied to the solution of the equation

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

   and has produced the results given below

```
n              x(n)                          g(x(n))
---------------------------------------------------------
```

```
0      2.100000000000000e-01      1.251610000000001e-01
1      2.391907083051520e-01     -2.904027288519462e-04
2      2.391232785547966e-01     -1.284441442095385e-09
3      2.391232782565544e-01      1.110223024625157e-16
4      2.391232782565545e-01                          0
```

Explain why you can not trust the computed values of $g(x_3)$ and $g(x_4)$.

Solution In the vicinity of a root, we will necessarily experience catastrophic cancellation, when we are computing g and add the last constant term. Therefore, as we converge, the computed values of $g(x_n)$ become increasingly unreliable. Finally, the statement $g(x_4) = 0$ indicates that the root is a rational number, which is somewhat unlikely considering solution formula for cubic equations.

3. (10 points) Find an interval of length at most 2×10^{-6} which is certain to contain the smallest positive root of g and determine the root with a relative error which is less than 10^{-6}.

   Solution We have an excellent candidate root, name $\xi = x_4 > 0.2$. With a view towards the desired relative error we consider the points $\xi \pm 10^{-7}$. We find
   $$g(\xi \pm 10^{-7}) = \mp 4.3067 \times 10^{-6}$$
   from which is follows that the interval $(\xi - 10^{-7}, \xi + 10^{-7})$ of length 2×10^{-7} is certain to contain a root r, and the midpoint ξ approximates the root with a relative error bounded by 10^{-6}, simply because
   $$\frac{|r - \xi|}{|r|} < \frac{10^{-7}}{0.2} = 5 \times 10^{-7} < 10^{-6}.$$

Problem 3 Let $\phi : \mathbb{R} \to \mathbb{R}$ be any function which is infinitely often differentiable.

1. (5 points) Let $x \in \mathbb{R}$ and let $h > 0$. Show that $D_h(x)$ given by

$$D_h(x) = \frac{\phi(x+h) - 2\phi(x) + \phi(x-h)}{h^2}$$

satisfies

$$D_h(x) = \phi''(x) + O(h^2).$$

Solution Let $x \in \mathbb{R}$ be given. Then by Taylor's formula there exist points ξ and ν such that

$$\phi(x+h) = \phi(x) + \phi'(x)h + \frac{\phi''(x)}{2}h^2 + \frac{\phi^{(3)}(x)}{3!}h^3 + \frac{\phi^{(4)}(\xi)}{4!}h^4$$

and

$$\phi(x-h) = \phi(x) - \phi'(x)h + \frac{\phi''(x)}{2}h^2 - \frac{\phi^{(3)}(x)}{3!}h^3 + \frac{\phi^{(4)}(\nu)}{4!}h^4$$

It follows that

$$\phi(x+h) + \phi(x-h) = 2\phi(x) + 2\frac{\phi''(x)}{2}h^2 + \frac{\phi^{(4)}(\xi) + \phi^{(4)}(\nu)}{4!}h^4$$

Therefore

$$D_h(x) = \phi''(x) + \frac{\phi^{(4)}(\xi) + \phi^{(4)}(\nu)}{4!}h^2$$

where the error term is $O(h^2)$ on each closed and bounded interval of $\mathbb{R}$.

2. (10 points) A specific ϕ has been chosen, together with the point $x_0 = 1$. Then $D_h(x_0)$ has been computed for $h = 2^{-k}x_0$, where $k = 1, 2, \ldots, 20$. The results can be found in the following table.

k	Dh	(Dh-D2h)	(D2h-D4h)/(Dh-D2h)
1	-2.629859253893		
2	-2.394779366424	2.350798874690e-01	
3	-2.314160132340	8.061923408471e-02	2.915928067761
4	-2.292594257216	2.156587512359e-02	3.738277886832
5	-2.287113989657	5.480267559562e-03	3.935186537738
6	-2.285738363786	1.375625870423e-03	3.983835777874
7	-2.285394109742	3.442540441938e-04	3.995961394281
8	-2.285308024504	8.608523785369e-05	3.998990451521
9	-2.285286501836	2.152266824851e-05	3.999747469028
10	-2.285281121149	5.380687071010e-06	3.999985125406

11	-2.285279775970	1.345179043710e-06	3.999978364344
12	-2.285279439762	3.362074494362e-07	4.001038781163
13	-2.285279363394	7.636845111847e-08	4.402439024390
14	-2.285279333591	2.980232238770e-08	2.562500000000
15	-2.285279273987	5.960464477539e-08	0.500000000000
16	-2.285279273987	0.000000000000e+00	Inf
17	-2.285280227661	-9.536743164062e-07	-0.000000000000
18	-2.285278320312	1.907348632812e-06	-0.500000000000
19	-2.285278320312	0.000000000000e+00	Inf
20	-2.285278320312	0.000000000000e+00	NaN

Determine the range of k for which we will be able to trust the corresponding error estimates.

Solution In view of the error expansion which we have just derived, the fractions must converge monotonically to 4 as h tends to zero. Inspecting the numbers, we find monotone convergence to 4 from below for $k = 3$ to $k = 10$. At $k = 11$ the upward trend is broken a bit. The value at $k = 12$ has jumped to the other side of 4, which shows that it is no longer correct to ignore the rounding errors. I would trust the sign, the magnitude as well as the first couple of digits of the error estimates for $k = 4, \dots, 10$, because the fractions are not only close to 4 but converging monotonically to 4 in this range of k.

Remark 1 The value at $k = 11$ is not bad, but it does not conform to the theoretical pattern, so if we want to play it safe, then we stop at $k = 10$.

3. (10 points) Find the value of $\phi''(1)$ with a relative error which is at most 10^{-6}.

   Solution We are not handed the error estimates directly, because the third column contains $D_h - D_{2h}$ rather than $(D_h - D_{2h})/3$. The number 3 is obtained by inspecting the fractions which are observed to converge towards 4 until rounding errors become a problem. It is clear that absolute value of $\phi''(1)$ is larger than 2. Therefore, dividing the numbers in the third column with 6, will give us a good relative error estimate!

   It is easy to see that $k = 10$ is the smallest integer that we can use and that this value falls in the range where we can trust the error estimate. We have
$$\phi''(1) \approx -2.285281121149.$$

Problem 4 Consider the problem of computing

$$f(\alpha) = \sqrt[5]{\alpha}$$

using a binary computer.

1. (5 points) Explain carefully why the problem is equivalent to solving the nonlinear equation

$$g(x) = 0, \quad \text{where} \quad g(x) = x^5 - \alpha, \tag{2}$$

and write down Newton's iteration for equation (2).

Solution We have

$$g(x) = 0 \Leftrightarrow x^5 = \alpha \Leftrightarrow x = \sqrt[5]{\alpha}$$

for the simple reason that $x \to x^5$ is strictly increasing for all $x > 0$. Hence there is an inverse function $x \to \sqrt[5]{x}$. This is why the last bi-implication is correct! The first bi-implication is trivial.

2. (5 points) Explain carefully why the problem is essentially solved if we can compute $f(x)$ for all machine numbers $x \in [1, 32]$.

Solution Any nonzero floating point number can be written in the form

$$x = (-1)^s (1.f)_2 \times 2^m$$

for some integer m. Now, $m = 5q + r$ (division with remainder) where $r \in \{0, 1, 2, 3, 4\}$. Therefore

$$\sqrt[5]{x} = (-1)^s \sqrt[5]{(1.f)_2 \times 2^r} \times 2^q$$

The only real problem is to compute $\sqrt[5]{(1.f)_2 \times 2^r}$. The numbers

$$(1.f)_2 \times 2^r,$$

all fall in the range $[1, 32)$.

3. (15 points) It is clear that we will need an intelligent way of initializing Newton's iteration, i.e. a function

$$x_0 = x_0(\alpha)$$

defined for $\alpha \in [1, 32]$. Assuming that we chosen a step-size $h > 0$ and have defined points

$$t_j = 1 + jh, \quad j = 0, 1, 2, \ldots, N, \quad Nh = 31,$$

and that we are willing to precompute the values

$$f(t_j), \quad j = 0, 1, 2, \ldots, N,$$

then we can define x_0 by interpolating f on each sub-interval $[t_j, t_{j+1}]$ using the corresponding first order polynomial. Now, what is the smallest value of N for which you will be able to ensure, that

$$|x_0(\alpha) - f(\alpha)| \leq \frac{1}{50}$$

for all $\alpha \in [1, 32]$?

Solution Let $I_h = [a, b]$ be any closed sub-interval of $[1, 32]$ of length $h > 0$. Let p be the polynomial of order 1 which interpolates f at the two endpoints. Let $x \in I_h$. Then there exists $\xi \in I_h$ such that

$$f(x) - p(x) = \frac{f^{(2)}(\xi)}{2}(x-a)(x-b)$$

Now, $f(x) = \sqrt[5]{x} = x^{\frac{1}{5}}$ and so $f'(x) = \frac{1}{5}x^{-\frac{4}{5}}$ and $f''(x) = -\frac{4}{25}x^{-\frac{9}{5}}$. It follows that

$$|f(x) - p(x)| \leq \frac{4}{25}\frac{1}{2}\frac{h^2}{4} \leq \frac{1}{50}$$

provided $h \leq 1$. Thus we can succeed with as little as $N = 32$ points and $h = 1$.