**Final #2 - Questions and Solutions**

1. Consider the following MATLAB code:

**while** $x > 9 * x/10$
  $x = 9 * x/10;$
**end**;
$x$

Suppose the above MATLAB code is run on a system working with the IEEE standard, double precision and round-to-nearest mode, and the initial value of $x$ lies within the interval $[1, 2]$. Explain why the above loop terminates its action within finitely many steps, and find the value of $x$ upon its termination.

The successive values $x$ attains form a descending sequence of floating point non-negative numbers. As there are only finitely many such numbers on the computer, the loop must eventually finish.

The final value of $x$ is a sub-normal number. In fact, the ratio between neighbouring normal numbers is about $1 + \varepsilon$. Hence for normal $y$ we certainly have $y > \text{round}(9 * y/10)$.

Each sub-normal number $x$ is of the form $x = ns$, where $s = 2^{-1074}$ is the smallest positive sub-normal number and $n$ is integer. The same reasoning as before shows that for relatively large $n$ we have $x > \text{round}(9 * x/10)$. More precisely, this is the case as long as $ns - s/2 > 9ns/10$. In other words, the value of $x$ is left unchanged if:

$$\frac{9ns}{10} > ns - \frac{s}{2} \ .$$

Equivalently:

$$9ns > 10ns - 5s \ ,$$

which yields:

$$n < 5 \ .$$

The case $n = 5$ is a borderline case. The exact value of $\frac{9 \cdot 5s}{10}$ is $4.5s$, which is equi-distant from its floating point neighbours, $4s$ and $5s$. Is is rounded to $4s$ since is such cases the rounding is to the number with a 0 digit at the lowest bit.

Thus, the loop will reduce the value of $x$ as long as $x > 4s$, and will be terminated at $x = 4s$.

Let us note that in case the initial value of $x$ is $s$, $2s$ or $3s$, the loop terminates immediately without changing the value of $x$. In our case, since the initial value of $x$ is big, the above calculations show that at no stage will it be rounded to one of these three values.

2. Let $f(x) = 2^x - 2x$. Show that $f$ has exactly two real roots. Find for which starting points, Newton's method leads to convergence to each of these roots.

Clearly, $x_1 = 1$ and $x_2 = 2$ are both roots of $f$. Now:

$$f'(x) = 2^x \ln 2 - 2 \,.$$

To find the zeros of $f'$ we solve $2^a \ln 2 - 2 = 0$, which gives $a = \log_2(\frac{2}{\ln 2}) = 1 - \frac{\ln \ln 2}{\ln 2}$. Note that $1 < a < 2$. Moreover, $f'(x) < 0$ for $x < a$ and $f'(x) > 0$ for $x > a$.

Now $f''(x) = 2^x \ln^2 2 > 0,\ x \in \mathbf{R}$, whence $f$ is convex on the whole line.

We cannot start Newton's method at the point $a$ (the tangent to the graph of $f$ is horizontal). Suppose $1 < x_0 < a$. The convexity of $f$ then gives $x_1 < 1$. Similarly, if $a < x_0 < 2$, then $x_1 > 2$. Hence, without loss of generality we may assume that the initial point does not belong to the interval $[1, 2]$.

Suppose first we start with a point $x_0 \in (-\infty, 1)$. The sequence $(x_n)_{n=0}^{\infty}$ is increasing and bounded above by 1 due to the convexity of $f$. Its convergence to the root 1 may be deduced from the general theorem proved in class or be proved directly as follows. Suppose $x_n \underset{n\to\infty}{\longrightarrow} b$. Then also $x_{n+1} \underset{n\to\infty}{\longrightarrow} b$, which means that

$$x_n - \frac{f(x_n)}{f'(x_n)} \underset{n\to\infty}{\longrightarrow} b.$$

Since $f$ and $f'$ are continuous, and $f'$ does not vanish to the left of 1, we obtain

$$x_n - \frac{f(x_n)}{f'(x_n)} \underset{n\to\infty}{\longrightarrow} b - \frac{f(b)}{f'(b)}.$$

Therefore $b - \frac{f(b)}{f'(b)} = b \quad \Rightarrow f(b) = 0 \quad \Rightarrow b = 1$.

In the same way one can easily prove that, starting from a point $x_0 > 2$, Newton's method gives a sequence converging to 2.

Summarizing the above, if we start at a point $x_0 < a$ we obtain a sequence converging to 1, while if we start at a point $x_0 > a$, the resulting sequence converges to 2.

3. Let $a = x_0 < x_1 < \ldots < x_n = b, \quad f_0 < f_1 < \ldots < f_n$. Prove or disprove the following statements:

a. The interpolation polynomial of degree not exceeding $n$ passing through the points $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n)$ forms an increasing function on the interval $[a, b]$.

False. For example, the parabola $y = -x^2$ is the interpolation polynomial corresponding to the points $(-2, -4), (-1, -1), (\frac{1}{2}, -\frac{1}{4})$, Although these points satisfy the conditions in question, the parabola does not increase on the whole interval $[-2, \frac{1}{2}]$.

b. The linear spline passing through those points forms an increasing function on the interval.

True. The equation of the spline in a typical sub-interval $[x_i, x_{i+1}]$ is

$$S(x) = f_i + \frac{f_{i+1} - f_i}{x_{i+1} - x_i}(x - x_i) ,$$

which is evidently increasing.

c. Every cubic spline passing through those points forms an increasing function on the interval.

False. When choosing a spline, we may take $S'(a)$ and $S'(b)$ arbitrarily. In particular, we may take $S'(a) < 0$, so that the spline will decrease in some neighbourhood of $a$.

4. We are looking for an approximation formula of the type

$$\int_{-1}^{1} f(x)dx \approx w_1 f(-1) + w_2 f(x_2) + \ldots + w_k f(x_k) ,$$

which will be exact for all polynomials up to some degree, as large as possible.

a. Explain intuitively for polynomials up to what degree is it plausible to expect such a formula to be precise.

We have $2k - 1$ free parameters. Hence we may expect to find a formula which will be exact for all polynomials of degree $\leq 2k - 2$.

b. Find $w_1, w_2, x_2$ for which the required formula is obtained in the case $k = 2$.

For the polynomails $1, x, x^2$, the following equalities are required:
(1) $\int_{-1}^{1} 1dx = 2 = w_1 \cdot 1 + w_2 \cdot 1$ ,
(2) $\int_{-1}^{1} xdx = 0 = w_1 \cdot (-1) + w_2 \cdot x_2$ ,
(3) $\int_{-1}^{1} x^2dx = \frac{2}{3} = w_1 \cdot (-1)^2 + w_2 \cdot x_2^2$ .
From (2) it follows that $w_1 = w_2x_2$. Substituting in (1) and (3) we obtain:
(4) $w_2(x_2 + 1) = 2$ ,
(5) $w_2x_2(x_2 + 1) = \frac{2}{3}$ .
Dividing both sides of (5) by the respective sides of (4) we get $x_2 = \frac{1}{3}$. From (4) it now follows that $w_2 = \frac{3}{2}$ and therefore $w_1 = \frac{1}{2}$.

c. For arbitrary fixed $k$, let $P(x) = (x - x_2) \cdot \ldots \cdot (x - x_k)$. Define a suitable inner product on the space of polynomials and explain how it enables in principle to find the polynomial $P$.

Define $\langle \cdot, \cdot \rangle$ by $\langle Q_1, Q_2 \rangle = \int_{-1}^{1}(x + 1)Q_1(x)Q_2(x)dx$ . The bilinearity and symmetry of $\langle \cdot, \cdot \rangle$ are straightforward. The inequality $\langle Q, Q \rangle > 0$ for $Q \neq 0$ follows from the fact that $x + 1$ is positive in the given interval (except for the point $-1$).

For the polynomial $P$ defined above and for any $Q$ of degree $\le k-2$ we have

$$\begin{aligned}\langle P, Q\rangle &= \int_{-1}^{1}(x+1)P(x)Q(x)dx \\ &= w_1 \cdot (-1+1)P(-1)Q(-1) + w_2 P(x_2)Q(x_2) \ldots w_k P(x_k)Q(x_k)\end{aligned}$$

(since the integrand is of degree $\le 2k-2$). Now the right hand side vanishes, the first term due to the factor $(-1+1)$ and all others since $P(x_i) = 0$ for $2 \le i \le k$.

Consequently, it is possible to find the polynomial $P$ for any $k$ using the Gram-Schmidt process.