Review Questions

Mark the correct answer in each part of the following questions.

1. We are working with a system implementing the IEEE standard with single precision and rounding to the nearest. Denote by $\odot$ and $\oslash$ the binary operations of multiplication and division, respectively, as peformed on floating point numbers in our system.

   (a) Let a_1, a_2 be positive normal numbers and s_1, s_2 positive subnormal numbers.

   (i) Each of the three relations $a_1 \odot a_2 > a_1 \cdot a_2$, $a_1 \odot a_2 = a_1 \cdot a_2$ and $a_1 \odot a_2 < a_1 \cdot a_2$ is possible. Similarly, each of the three relations $s_1 \odot s_2 > s_1 \cdot s_2$, $s_1 \odot s_2 = s_1 \cdot s_2$ and $s_1 \odot s_2 < s_1 \cdot s_2$ is possible.

   (ii) Each of the above three relations invloving the a_i's is possible, but only two of those invloving the s_i's are possible.

   (iii) Each of the above three relations invloving the a_i's is possible, but only one of those invloving the s_i's is possible.

   (iv) Only two of the above three relations invloving the a_i's are possible, and only two of those invloving the s_i's are possible.

   (v) None of the above.

   (b) The sum of all positive normal numbers (i.e., the actual sum, not the sum calculated by the system) is:

   (i) $3 \cdot 2^{150} - 2^{127} - 3 \cdot 2^{-104} + 2^{-127}$.

   (ii) $9 \cdot 2^{150} - 2^{127} - 9 \cdot 2^{-104} + 2^{-127}$.

   (iii) $3 \cdot 2^{150} + 2^{127} + 3 \cdot 2^{-104} + 2^{-127}$.

   (iv) $9 \cdot 2^{150} + 2^{127} + 9 \cdot 2^{-104} + 2^{-127}$.

   (v) None of the above.

(c) Consider the following three possible properties of a floating point number a in the interval $[1, 2)$:

A. $a \oslash 3 > a/3$.

B. $a \oslash 3 = a/3$.

C. $a \oslash 3 < a/3$.

(i) There exist numbers satisfying Property A, there exist numbers satisfying Property B, and there exist numbers satisfying Property C.

(ii) All numbers satisfy Property B.

(iii) There exist numbers satisfying Property A, there exist numbers satisfying Property B, but there exist no numbers satisfying Property C.

(iv) There exist no numbers satisfying Property A, but there exist numbers satisfying Property B and there exist numbers satisfying Property C.

(v) None of the above.

(d) Consider the Matlab code section

```
a=1;
while(a+eps>a)
   a=a+eps;
end;
a
```

We run the code on a system with the sepcifications listed at the beginning of the question. The output of this code is:

(i) 2.

(ii) The largest floating point number in the system.

(iii) ∞.

(iv) `NaN`.

(v) None of the above.

2. In this question we deal with fixed points of certain functions g. We start at some point x_0 and continue according to the iteration $x_{n+1} = g(x_n)$ for $n \geq 0$.

(a) The point 0 is a fixed point of both functions g_1 and g_2, defined by:

$$g_1(x) = \frac{1}{4}\sin x + \frac{3}{4}\operatorname{tg} x, \qquad g_2(x) = \frac{3}{4}\sin x + \frac{1}{4}\operatorname{tg} x.$$

(Hint: You may use the expansions

$$\begin{aligned} \sin x &= x - \tfrac{x^3}{6} + O\left(x^5\right), \\ \operatorname{tg} x &= x + \tfrac{x^3}{3} + O\left(x^5\right) \end{aligned}$$

of $\sin x$ and $\operatorname{tg} x$ near 0.)

(i) If x_0 is sufficiently close to 0 then the sequence $(x_n)_{n=1}^{\infty}$ corresponding to g_2 converges to 0, but the analogous sequence for g_1 does not. However, the convergence for g_2 is slower than linear.

(ii) If x_0 is sufficiently close to 0 then the obtained sequences converge to 0 for both g_1 and g_2. However, the convergence for g_1 is slower than linear, while for g_2 it is linear.

(iii) If x_0 is sufficiently close to 0 then the obtained sequences converge to 0 for both g_1 and g_2. The convergence is linear for g_1 and quadratic for g_2.

(iv) If x_0 is sufficiently close to 0 then the obtained sequences converge to 0 quadratically for both g_1 and g_2.

(v) None of the above.

(b) Let $g(x) = x^2 \cos x$. Notice that $\xi_0 = 0$ is a fixed point of g. In addition, the function has a fixed point $\xi_k \in (2k\pi, 2k\pi + \pi/2)$ for every positive integer k.

(i) For each $k \geq 0$, there exists no neighborhood U_k of ξ_k such that, if $x_0 \in U_k$, then $x_n \xrightarrow[n\to\infty]{} \xi_k$.

(ii) There exists a neighborhood U_0 of ξ_0 such that, if $x_0 \in U_0$, then $x_n \xrightarrow[n\to\infty]{} \xi_0$, where the convergence is linear. However, if k is sufficiently large, then no such neighborhood U_k exists for ξ_k.

(iii) There exists a neighborhood U_0 of ξ_0 such that, if $x_0 \in U_0$, then $x_n \xrightarrow[n\to\infty]{} \xi_0$, where the convergence is quadratic. However, if k is sufficiently large, then no such neighborhood U_k exists for ξ_k.

(iv) For each $k \geq 0$, there exists a neighborhood U_k of ξ_k such that, if $x_0 \in U_k$, then $x_n \underset{n\to\infty}{\longrightarrow} \xi_k$. However, whereas the convergence is quadratic for $k = 0$, it is only linear for $k \geq 1$.

(v) None of the above.

(c) The function $g : [2, 3] \longrightarrow [2, 3]$ is not necessarily continuous, yet is known to have a fixed point ξ. Consider the fixed point ξ^2 of the function $g_1 : [4, 9] \longrightarrow [4, 9]$, defined by:

$$g_1(x) = g\left(\sqrt{x}\right)^2, \qquad x \in [4, 9].$$

Consider the following possible properties of the functions:

A. There exists a neighborhood U of ξ such that, if $x_0 \in U$, then the sequence of iterates (x_n) under g satisfies $x_n \underset{n\to\infty}{\longrightarrow} \xi$, where the convergence is at least linear.

B. There exists a neighborhood U of ξ^2 such that, if $x_0 \in U$, then the sequence of iterates (x_n) under g_1 satisfies $x_n \underset{n\to\infty}{\longrightarrow} \xi^2$, where the convergence is at least linear.

(i) Property A is equivalent to Property B.

(ii) Property A implies Property B, but not vice versa.

(iii) Property B implies Property A, but not vice versa.

(iv) Neither property implies the other.

3. In this question we deal with zeros of certain functions f.

(a) Consider the functions f_1 and f_2, defined by:

$$f_1(x) = \ln(x^2 - 3), \qquad f_2(x) = x^2 - 4.$$

We are interested in the performance of Newton's method when trying to find the zeros of the two functions.

(i) Newton's method works equally well for the two functions. Namely, if when starting the iteration for f_1 from some point x_0 we converge to some zero of f_1 at some speed, then the same holds for f_2, and vice versa. Moreover, when starting from a point sufficiently close to one of the zeros, we converge quadratically to that zero.

(ii) Newton's method works well for both functions in the sense that each zero has a neighborhood such that, when starting from a point in this neighborhood, we converge to 0 for each of the functions. However, the speed of convergence when starting at such initial points is linear for one of the functions and quadratic for the other.

(iii) Newton's method converges for f_1 when starting from a point sufficiently close to one of the zeros. However, there are many starting points for which the method does not lead to a converging sequence. On the other hand, for f_2 there exists only one starting point on the real line for which we do not obtain a sequence converging to one of the zeros.

(iv) One of the two functions has the property that there exists points arbitrarily close to one of the zeros such that, when starting the iteration for this functions at one of these points, we converge to another zero of that function.

(v) None of the above.

(b) Let $f : (0, \infty) \longrightarrow \mathbf{R}$ be defined by

$$f(x) = x^{\alpha} e^{x} - e, \qquad x > 0,$$

where α is an arbitrary fixed real positive number.

(i) For every α and every initial point $x_0 > 0$ (where x_0 is not a zero of f), Newton's method converges at most linearly fast to a zero of f.

(ii) For every sufficiently large α, Newton's method converges quadratically when started in a sufficiently small neighborhood of the zero of f. However, for every α there exist initial values $x_0 > 0$ for which the method fails to converge to the zero of f, and there exist values of α for which there exists no neighborhood as above.

(iii) For every α, Newton's method converges quadratically when started in a sufficiently small neighborhood of the zero of f. However, for every α there exist initial values $x_0 > 0$ for which the method fails to converge to the zero of f.

(iv) For every α and $x_0 > 0$, Newton's method converges quadratically.

(v) None of the above.

(c) Consider the equation

$$e^x \arcsin x - 2x = 0,$$

which is equivalent to the fixed-point equation $g(x) = x$, where

$$g(x) = e^x \arcsin x - x, \qquad x \in [-1, 1].$$

The equation has two zeros $\xi_1 = 0$ and $\xi_2 \in [0.6, 0.7]$. We try to solve the equation by iterating g.

(i) The point ξ_1 has a neighborhood U such that, starting the iterations at a point $x_0 \in U$, we converge to ξ_1. The convergence in this case is roughly at the same speed as that of the bisection method. The point ξ_2 has no such neighborhood.

(ii) The point ξ_1 has a neighborhood U such that, starting the iterations at a point $x_0 \in U$, we converge to ξ_1 quadratically. The point ξ_2 has no neighborhood that guarantees convergence.

(iii) The point ξ_1 has a neighborhood U such that, starting the iterations at a point $x_0 \in U$, we converge to ξ_1 quadratically. The point ξ_2 has a neighborhood for which the same holds, but only linearly fast.

(iv) Both points $\xi_i, i = 1, 2$, have neighborhoods U_i such that, starting the iterations at a point $x_0 \in U_i$, we converge to ξ_i quadratically.

(v) None of the above.

Solutions

1. (a) Since $1 \cdot 1 = 1$ is a floating point number, we have $1 \odot 1 = 1 \cdot 1$.

   We have $3 \cdot (1 + 2^{-23}) = 2 + 1 + 2^{-22} + 2^{-23}$. Normalizing, we obtain the representation $(1 + 2^{-1} + 2^{-23} + 2^{-24}) \cdot 2^1$, which needs to be up rounded. Thus $3 \odot (1 + 2^{-23}) > 3 \cdot (1 + 2^{-23})$.

   Now $(1 + 2^{-23}) \cdot (1 + 2^{-23}) = 1 + 2^{-22} + 2^{-46}$, which needs to be down rounded to $1 + 2^{-22}$. Thus $(1 + 2^{-23}) \odot (1 + 2^{-23}) < (1 + 2^{-23}) \cdot (1 + 2^{-23})$.

   Altogether, all 3 orderings are possible between $a_1 \odot a_2$ and $a_1 \cdot a_2$.

   Clearly, $s_1 \cdot s_2 < 2^{-126} \cdot 2^{-126} = 2^{-252}$, which needs to be down rounded to 0. Hence we necessarily have $s_1 \odot s_2 < s_1 \cdot s_2$.

   Thus, (iii) is true.

   (b) Denote by M the set of all floating point numbers in the interval $[1, 2)$ and by T the set of all powers of 2 from 2^{-126} up to 2^{127}. The required sum S is

$$\sum_{m \in M, t \in T} mt = \sum_{m \in M} m \cdot \sum_{t \in T} t. \tag{1}$$

   The first factor on the right-hand side is the sum of an arithmetic progression, whose first term is 1, whose last term is $2 - 2^{-23}$, and whose length is 2^{23}. Hence:

$$\sum_{m \in M} m = (1 + 2 - 2^{-23}) \cdot 2^{23}/2 = 3 \cdot 2^{22} - 2^{-1}.$$

   The second factor on the right-hand side of (1) is a sum of a geometric progression, so that:

$$\sum_{t \in T} t = 2^{128} - 2^{-126}.$$

   It follows that:

$$S = (3 \cdot 2^{22} - 2^{-1}) \cdot (2^{128} - 2^{-126}) = 3 \cdot 2^{150} - 2^{127} - 3 \cdot 2^{104} + 2^{-127}.$$

   Thus, (i) is true.

(c) $(3/2)/3 = 1/2$ is a floating point number, so that the number $3/2$ satisfies B.

The infinite binary expansion of the number $1/3$ is $0.0101...$, which may be written in the form $1.0101... \cdot 2^{-2}$, and is therefore up rounded to $1.0101...01011 \cdot 2^{-2}$ in our system. Thus, the number 1 satisfies A.

Now

$$(7/4)/3 = 1/2 + 1/12 = (1 + 1/6) \cdot 2^{-1} = \left(1 + \sum_{n=1}^{\infty} 2^{-2n-1}\right) \cdot 2^{-1},$$

which is down rounded to $(1 + 2^{-3} + 2^{-5} + 2^{-7} + \ldots + 2^{-23}) \cdot 2^{-1}$. It follows that the number $7/4$ satisfies C.

Thus, (i) is true.

(d) The floating point numbers between 1 and 2 are the numbers $1, 1+\varepsilon, 1+2\varepsilon, \ldots, 2$. Thus the loop will change a from 1 to 2 within 2^{23} steps. Now $(2+\varepsilon)_+ = 2 + 2\varepsilon$, while $(2+\varepsilon)_- = 2$, so that $2 \oplus \varepsilon = 2$. Hence the loop will stop when a becomes 2.

Thus, (i) is true.

2. (a) We have

$$g_1'(x) = \frac{1}{4}\cos x + \frac{3}{4\cos^2 x}, \qquad g_2'(x) = \frac{3}{4}\cos x + \frac{1}{4\cos^2 x}.$$

Hence:

$$g_1'(0) = g_2'(0) = 1.$$

Thus, for both functions we are in a borderline case; we may have divergence, but we may also have (slow) convergence. To decide, we need to consider the functions g_1 and g_2 more carefully. Near 0 we have

$$g_1(x) = x + \frac{5}{24}x^3 + O\left(x^5\right)$$

and

$$g_2(x) = x - \frac{1}{24}x^3 + O\left(x^5\right).$$

It follows that, if x is sufficiently close to 0, then $g_2(x)$ is slightly closer to 0 than is x, while $g_1(x)$ is slightly farther. Consequently, for g_1 we certainly do not have convergence. Since we can find a neighborhood of 0 not including any fixed point of g_2 but 0, this also means that for g_2 the sequence does converge to 0 if we start sufficiently close to 0. More accurately, from the above we see that the sequence of errors (e_n) satisfies $e_{n+1} \approx e_n - e_n^3/24$.

Thus, (i) is true.

(b) We have

$$g'(x) = 2x \cos x - x^2 \sin x,$$

so that $g'(0) = 0$ and the convergence is quadratic. To obtain a more precise estimate we calculate

$$g''(x) = 2 \cos x - 4x \sin x - x^2 \cos x,$$

so that $g''(0) = 2$ and $e_{n+1} \approx -e_n^2$. (Of course, we have exactly $x_{n+1} = x_n^2 \cos x_n$, which gives $e_{n+1} = -e_n^2 \cos e_n$, yielding the above estimate.)

Now take an arbitrary fixed $k \geq 1$, and put $\xi = \xi_k$. Employing the equality $g(\xi) = \xi$, we obtain $\xi \cos \xi = 1$, and therefore:

$$\begin{aligned} g'(\xi) &= 2\xi \cos \xi - \xi^2 \sin \xi \\ &= 2 - \xi^2 \sqrt{1 - 1/\xi^2} \\ &= 2 - \xi \sqrt{\xi^2 - 1} \\ &< 2 - 2\pi \sqrt{4\pi^2 - 1} \\ &< 2 - 6 \cdot 6 = -34. \end{aligned}$$

Thus, ξ_k has no neighborhood guaranteeing convergence.

Thus, (iii) is true.

(c) We claim that A and B are equivalent. In fact, suppose first that A is satisfied for some neighborhood U of ξ. Consider the neighborhood $U^2 = \{x^2 : x \in U\}$ of ξ^2. Let $x_{01} \in U^2$ be a starting point for the iterations for g_1. Let $(x_{n1})_{n=0}^{\infty}$ the resulting sequence of iterates for g_1. Starting to iterate for g at the point $x_0 = \sqrt{x_{01}}$, we easily show by induction that we obtain the sequence $(x_n)_{n=0}^{\infty}$ with $x_n = \sqrt{x_{n1}}$ for each n.

Since A is satisfied for U, we have $|\xi - x_{n+1}| \leq \alpha|\xi - x_n|$ for all sufficiently large n, where $\alpha < 1$. Hence, applying the mean-value theorem to the mapping $t \mapsto t^2$, we obtain:

$$|\xi - x_{n+1,1}| = |\xi^2 - x_{n+1}^2| = |2\eta| \cdot |\xi - x_{n+1}| \leq |2\eta| \cdot \alpha \cdot |\xi - x_n|,$$

where η is an intermediate point between ξ and x_{n+1}. (Of course, there is no need to invoke the mean-value theorem here, as the equality clearly holds with $\eta = (\xi + x_{n+1})/2$. However, it shows that if we had any differentiable function instead of the square function, it would work just as well. This is what needs to be done later for proving the inverse direction.) By the same token

$$|\xi - x_{n,1}| = |2\eta'| \cdot |\xi - x_n|,$$

where η' lies between ξ and x_n, and therefore:

$$|\xi - x_n| = \frac{1}{|2\eta'|} \cdot |\xi - x_{n,1}|.$$

It follows that:

$$|\xi - x_{n+1,1}| \leq \left|\frac{\eta}{\eta'}\right| \cdot \alpha \cdot |\xi - x_{n,1}|.$$

Since both η and η' lie between ξ and x_n (or x_{n+1}), the ratio η/η' becomes arbitrarily close to 1 as $n \to \infty$. Taking $\alpha' > \alpha$ (but still $\alpha' < 1$), we get

$$|\xi - x_{n+1,1}| \leq \alpha'|\xi - x_{n,1}|$$

for all sufficiently large n. Hence A implies B.

The inverse direction works in the same way, with the square function replaced by the square root function.

Thus, (i) is true.

3. (a) First note that the only zeros of both f_1 and f_2 are 2 and -2. Given any even function, Newton's method works for it the same way when started from a point $x_0 > 0$ as from $-x_0$. Hence for both functions we will deal only with convergence to the zero $\xi = 2$.

(Note that f_1 is undefined at 0 and $f_2'(0) = 0$, so we will not start from $x_0 = 0$.)

Since $f_2'(\xi) = 4$, Newton's method converges at least quadratically for f_2 when started in a sufficiently small neighborhood of ξ. We claim that the same holds for any starting point $x_0 > 0$. In fact, note that f_2 is increasing and convex throughout $(0, \infty)$. Hence the method certainly converges when $x_0 > 2$. If $0 < x_0 < 2$, then clearly $x_1 > 2$, and again we have convergence.

Newton's method converges at least quadratically for f_1 when started in a sufficiently small neighborhood of ξ for the same reason, namely that $f_1'(\xi) = 4$. However, the situation is different if we start farther farther away from ξ. Note first that f is increasing and concave throughout its domain of definition in the positive axis, namely $(\sqrt{3}, \infty)$. Hence Newton's method converges if $\sqrt{3} < x_0 < \xi = 2$. Now suppose $x_0 > \xi$. Obviously, as x_0 increases, x_1 decreases. Moreover, since $f_1(x) \underset{x\to\infty}{\longrightarrow} \infty$ and $f_1'(x) \underset{x\to\infty}{\longrightarrow} 0$, as x_0 increases from ξ to ∞, the point x_1 decreases continuously from ξ to $-\infty$. In particular, there is some interval $I \subseteq (\xi, \infty)$ such that if $x_0 \in I$ then $x_1 \in [-\sqrt{3}, \sqrt{3}]$, namely Newton's method is stuck after a single iteration.

Thus, (iii) is true.

(b) Since f is increasing throughout its domain of definition and $f(1) = 0$, the point $\xi = 1$ is the only zero of f. We have $f'(1) = (\alpha + 1)e > 0$, so that Newton's method converges at least quadratically if we start with x_0 sufficiently close to ξ.

We have

$$f''(x) = \left(x^\alpha + 2\alpha x^{\alpha-1} + \alpha(\alpha - 1)x^{\alpha-2}\right) e^x.$$

Since $f''(1) \neq 0$, the convergence is quadratic. A routine calculation shows that, for $\alpha \geq 1$, the function is convex throughout $(0, \infty)$; for $\alpha < 1$, the function is concave on $(0, -\alpha + \sqrt{\alpha})$ and convex thereafter. Hence if $\alpha \geq 1$ then Newton's method converges monotonically if $x_0 > \xi$, while if $x_0 < \xi$ then $x_1 > \xi$ and thereafter we have monotonic convergence. Now suppose $\alpha < 1$. Again, starting to the right of ξ we get monotonic convergence, while starting anywhere in the interval $[-\alpha + \sqrt{\alpha}, \xi)$ we get to a point x_1 to the right of ξ and have monotonic convergence thereafter. If $x_0 < -\alpha + \sqrt{\alpha}$, then we will get consecutively larger $x_1, x_2, \ldots$, until some x_k will already be to the right of $-\alpha + \sqrt{\alpha}$ (either to the right or to the left of ξ) and we are back to the former case.

Thus, (iv) is true.

(c) We have

$$g'(x) = e^x \arcsin x + \frac{e^x}{\sqrt{1-x^2}} - 1.$$

At ξ_1 the function g' vanishes, so by iterating g, starting at a sufficiently small neighborhood of ξ_1, we have quadratic convergence to ξ_1. Now:

$$\begin{aligned} g'(\xi_2) &= e^{\xi_2} \arcsin \xi_2 + \frac{e^{\xi_2}}{\sqrt{1-\xi_2^2}} - 1 \\ &= 2\xi_2 + \frac{e^{\xi_2}}{\sqrt{1-\xi_2^2}} - 1 \\ &\geq 2 \cdot 0.6 + \frac{e^{0.6}}{\sqrt{1-0.6^2}} - 1 \\ &\geq 1.2 + \frac{1}{\sqrt{1-0.6^2}} - 1 = 1.45. \end{aligned}$$

Hence we do not obtain convergence when starting near ξ_2.

Thus, (ii) is true.