Midterm

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 $\oplus$ the binary operation of addition, as performed on floating point numbers in our system, and denote analogous operations similarly.

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

   (i) We necessarily have $a_1 \oplus a_2 > a_1 \oplus s_2 > s_1 \oplus s_2 > s_1$.

   (ii) We necessarily have $a_1 \oplus a_2 > a_1 \oplus s_2$ and $s_1 \oplus s_2 > s_1$, but may have $a_1 \oplus s_2 = s_1 \oplus s_2$.

   (iii) We may have $a_1 \oplus a_2 = a_1 \oplus s_2$ and $a_1 \oplus s_2 = s_1 \oplus s_2$, but we necessarily have $s_1 \oplus s_2 > s_1$.

   (iv) We may have $a_1 \oplus a_2 = a_1 \oplus s_2$, but we necessarily have $a_1 \oplus s_2 > s_1 \oplus s_2 > s_1$.

   (v) None of the above.

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

   (i) $(2^{23})!/2^{149\cdot(2^{23}-1)}$.

   (ii) $(2^{23}-1)!/2^{149\cdot 23}$.

   (iii) $(2^{23})!/2^{149\cdot 23}$.

   (iv) $(2^{23}-1)!/2^{149\cdot(2^{23}-1)}$.

   (v) None of the above.

   (c) Consider the equations $(2 \oslash 3) \odot x = x$ and $(3 \oslash 2) \odot x = x$ in positive floating point numbers.

   (i) Both equations have no solutions.

(ii) Both equations have a unique solution.

(iii) The first equation has no solutions while the second has exactly one.

(iv) The first equation has exactly one solution while the second has none.

(v) None of the above.

(d) Consider the Matlab code section

```
b=a; for i=1:k a=2*a; b=0.500001*b; end; a/b
```

where a is some positive floating point number and k some positive integer, both defined earlier. We run the code on a system with the specifications listed at the beginning of the question.

(i) For $k = 2$ there exists a value of a for which the output of the above section is 12. For $k = 130$ the output must be ∞.

(ii) For $k = 2$ the output of the above section is either 16 or ∞. For $k = 130$ the output must be ∞.

(iii) For $k = 2$ there exists a value of a for which the output of the above section is 12. For $k = 130$ the output may be ∞, but may also be a finite floating point number.

(iv) For $k = 2$ the output of the above section is either 16 or ∞. For $k = 130$ the output may be ∞, but may also be a finite floating point number.

(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) Let:

$$g(x) = \begin{cases} x^{2/3}, & x \geq 0, \\ \\ -|x|^{2/3}, & x < 0. \end{cases}$$

Notice that g has exactly 3 fixed points, namely $\xi_1 = -1, \xi_2 = 0, \xi_3 = 1$.

(i) For every choice of $x_0 \in \mathbf{R}$, the sequence $(x_n)_{n=1}^{\infty}$ converges to one of the fixed points. Moreover, each of the fixed points ξ_i has a neighborhood U_i, such that, if $x_0 \in U_i$, then $x_n \underset{n\to\infty}{\longrightarrow} \xi_i$.

(ii) For every choice of $x_0 \in \mathbf{R}$, the sequence $(x_n)_{n=1}^{\infty}$ converges to one of the fixed points. However, some of the fixed points ξ_i have no neighborhood U_i as in (i).

(iii) There exist choices of x_0 for which the sequence $(x_n)_{n=1}^{\infty}$ diverges. However, each of the fixed points ξ_i has a neighborhood U_i as in (i).

(iv) There exist choices of x_0 for which the sequence $(x_n)_{n=1}^{\infty}$ diverges. Moreover, some of the fixed points ξ_i have no neighborhood U_i as in (i).

(v) None of the above.

(b) Consider the function $g(x) = \operatorname{tg} x$. Notice that it has exactly one fixed point ξ_k in each interval of the form $(k\pi - \pi/2, k\pi + \pi/2)$ for integer k.

(i) For every k, the point ξ_k has a neighborhood U_k such that, if $x_0 \in U_k$, then $x_n \underset{n\to\infty}{\longrightarrow} \xi_k$.

(ii) There exist finitely many (but not 0) indices k for which ξ_k has no neighborhood U_k as in (i), but for all other k's there exists such a neighborhood.

(iii) There exist infinitely many indices k for which ξ_k has no neighborhood U_k as in (i), and infinitely many for which ξ_k does have such a neighborhood.

(iv) There exists exactly one index k for which ξ_k has a neighborhood U_k as in (i).

(v) None of the above.

(c) Suppose the point 4 is a fixed point of g, and:

$$g'(4) = g''(4) = 0, \qquad g'''(4) = -3.$$

We have started the iteration at some point x_0, and after 5 steps we are at $x_5 = 4 - 10^{-6}$. It is reasonable to guess that x_6 is approximately

(i) $4 - 5 \cdot 10^{-19}$.

(ii) $4 - 3 \cdot 10^{-12}$.
(iii) $4 + 3 \cdot 10^{-12}$.
(iv) $4 + 5 \cdot 10^{-19}$.
(v) None of the above.

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

   (a) The equation $x^3 - x = 0$ has the three zeros $\xi_1 = -1, \xi_2 = 0, \xi_3 = 1$. We employ Newton's method to solve the equation, starting from a certain point x_0.

   (i) Each of the zeros $\xi_i, 1 \le i \le 3$, has a neighborhood U_i such that, if $x_0 \in U_i$, then the resulting sequence converges at least quadratically to ξ_i. Moreover, $U_3 \supseteq (1, \infty)$.
   (ii) The zeros ξ_1 and ξ_3 have neighborhoods U_1 and U_3, respectively, such that, if $x_0 \in U_i$, then the resulting sequence converges at least quadratically to ξ_i, but ξ_2 has no such neighborhood. Moreover, $U_3 \supseteq (1, \infty)$.
   (iii) Each of the zeros $\xi_i, 1 \le i \le 3$, has a neighborhood U_i such that, if $x_0 \in U_i$, then the resulting sequence converges at least quadratically to ξ_i. However, $U_3 \not\supseteq (1, \infty)$.
   (iv) The zeros ξ_1 and ξ_3 have neighborhoods U_1 and U_3, respectively, such that, if $x_0 \in U_i$, then the resulting sequence converges at least quadratically to ξ_i, but ξ_2 has no such neighborhood. Also, $U_3 \not\supseteq (1, \infty)$.
   (v) None of the above.

   (b) Consider the equation

$$e^x = x^2 + 2x.$$

   Notice that the equation is equivalent to each of the equations $g_i(x) = x, i = 1, 2$, where the functions g_1, g_2 are defined by:

$$g_1(x) = \frac{e^x - x^2}{2}, \qquad g_2(x) = \ln\left(x^2 + 2x\right).$$

   Notice also that the difference $e^x - (x^2 + 2x)$ assumes values of opposite signs at the points 2.2 and 2.3, so that our equation has a solution $\xi \in [2.2, 2.3]$.

(i) Trying to solve the equation by iterating either g_1 or g_2, starting from a point sufficiently close to ξ, we obtain a sequence getting away from it.

(ii) Trying to solve the equation by iterating g_1, starting from a point sufficiently close to ξ, we obtain a sequence getting away from it. Trying to solve the equation by iterating g_2, starting from a point sufficiently close to ξ, we obtain a sequence converging linearly to ξ, which is much slower than does Newton's method in this case.

(iii) Trying to solve the equation by iterating either g_1 or g_2, starting from a point sufficiently close to ξ, we obtain a sequence converging linearly to ξ, which is much slower than does Newton's method in this case.

(iv) Trying to solve the equation by iterating g_1, starting from a point sufficiently close to ξ, we obtain a sequence converging linearly to ξ. Trying to solve the equation by iterating g_2, starting from a point sufficiently close to ξ, we obtain a sequence converging quadratically to ξ, which is roughly the speed provided by Newton's method in this case.

(v) None of the above.

(c) Consider the function f defined by:

$$f(x) = \begin{cases} -|\sin x|^{5/2}, & -\frac{\pi}{4} \leq x < 0, \\ (\sin x)^{5/2}, & 0 \leq x \leq \frac{\pi}{4}. \end{cases}$$

We solve the equation $f(x) = 0$ by Newton's method.

(i) The convergence is linear, but slightly slower than that of the bisection method.

(ii) The convergence is linear, with speed almost the same as that of the bisection method.

(iii) The convergence is linear, but slightly faster than that of the bisection method.

(iv) The convergence is quadratic.

(v) None of the above.

Solutions

1. (a) Obviously, the operation $\oplus$ is non-decreasing as a function of each of the variables. The question is to what extent it is strictly increasing. We check each inequality separately.

   - $a_1 \oplus a_2 > a_1 \oplus s_2$:
     Take $a_1 = 2^{100}, a_2 = 1$, and s_2 as any sub-normal number. Then:
     $$a_1 \oplus a_2 = a_1 = a_1 \oplus s_2.$$
   - $a_1 \oplus s_2 > s_1 \oplus s_2$:
     We claim that the inequality indeed holds. Since $a_1 \oplus s_2 \geq 2^{-126} \oplus s_2$ and $s_1 \oplus s_2 \leq (1 - 2^{-23}) \cdot 2^{-126} \oplus s_2$, it suffices to show that $2^{-126} \oplus s_2 > (1 - 2^{-23}) \cdot 2^{-126} \oplus s_2$. Now notice that all integer multiples of 2^{-149}, from $1 \cdot 2^{-149}$ up to $2^{24} \cdot 2^{-149}$, are floating point numbers. Therefore $2^{-126} \oplus s_2 = 2^{-126} + s_2$ and $(1 - 2^{-23}) \cdot 2^{-126} \oplus s_2 = (1 - 2^{-23}) \cdot 2^{-126} + s_2$. This proves the required inequality.
   - $s_1 \oplus s_2 > s_1$:
     As with the preceding part, we necessarily have $s_1 \oplus s_2 = s_1 + s_2 > s_1$.

   Thus, (iv) is true.

   (b) The set of positive sub-normal numbers is the set of all integer multiples of 2^{-149}, from $1 \cdot 2^{-149}$ up to $(2^{23} - 1) \cdot 2^{-149}$. Consequently, the required product is
   $$\prod_{i=1}^{2^{23}-1} \left(i \cdot 2^{-149}\right) = \frac{(2^{23} - 1)!}{2^{149(2^{23}-1)}}.$$
   Thus, (iv) is true.

   (c) Let $x = m \cdot 2^E$, where either $m = 1.b_1 b_2 \ldots b_{23}$ and $-126 \leq E \leq 127$ or $m = 0.b_1 b_2 \ldots b_{23}$ and $E = -126$. Unless $m = 0.00\ldots01$ and $E = -126$, we have $(3/2)x \geq (m + 2^{-23}) \cdot 2^E$, and since the right-hand side is a floating point number (or ∞) in our system we get $(3 \oslash 2) \odot x > x$. In the exceptional case where $m = 0.00\ldots01$ and $E = -126$, we verify that (due to the rounding rules) $(3 \oslash 2) \odot x =$

$0.00\ldots010{\cdot}2^{-126} = 2^{-148} > 2^{-149} = x$. Hence the second equation has no solution.

The situation regarding the first equation is similar, but here we need to check directly the two cases (i) $m = 0.00\ldots01$ and $E = -126$, and (ii) $m = 0.00\ldots010$. In the first of these, the equation is satisfied, while in the second it is not. Hence the equation has a unique solution.

Thus, (iv) is true.

(d) For $k = 2$, if we start with $a = 3 \cdot 2^{-149}$, then at the end of the execution of the program we have $a = 12 \cdot 2^{-149}$ and $b = 2^{-149}$, so that the output is 12.

For $k = 130$, in principle, since a is doubled at each iteration and b about halved, the final value of a/b is about 2^{260}, which is ∞ in our system. However, this is not completely correct, as b may become 2^{-149} at some point during the execution of the loop and stay so for the rest of the loop. Yet, even in this case, a will either double at each iteration, or become ∞ at some point and stay so for the rest of the loop. Hence, in any case, either a will be ∞ or it will be at least 2^{130} times its initial value, so that the output will be ∞.

Thus, (i) is true.

2. (a) Since g is an odd function, it suffices to understand it on $[0, \infty)$. We have $g'(x) = \frac{2}{3\sqrt[3]{x}} > 0$ for $x > 0$, and g' is undefined at 0. Obviously $0 \le g'(x) \le \frac{2}{3}$ for $x \ge 1$. Therefore, for $x_0 \in [1, \infty)$ the sequence $(x_n)_{n=1}^{\infty}$ decreases to ξ_3. For $x_0 \in (0, 1)$ the sequence is easily seen by induction to be increasing and bounded above by 1. Hence it must converge to a fixed point of g, which must be ξ_3.

The situation on $(-\infty, 0)$ is analogous. In particular, ξ_2 does not admit a neighborhood U_2 as required. However, for every $x_0 \in (0, \infty)$ the sequence $(x_n)_{n=1}^{\infty}$ converges to $\xi_3 = 1$ and for every $x_0 \in (-\infty, 0)$ it converges to $\xi_1 = -1$.

Thus, (ii) is true.

(b) Since $|g'(x)| = 1/\cos^2 x \ge 1$ at any point where g is defined, if $x \in (k\pi - \pi/2, k\pi + \pi/2)$ then

$$|\operatorname{tg} x - \xi_k| = \frac{1}{\cos^2 \eta} \cdot |x - \xi_k| \ge |x - \xi_k|, \qquad (\eta \in (\xi_k, x)).$$

Hence for no k can ξ_k have a neighborhood U_k as required.

(In fact, a point x_0 may lead to a fixed point under the iteration process only if some x_n happens to coincide with some ξ_k. Since the function g is countable-to-one, there may be only countably many points x_0 possessing this property.)

Thus, (v) is true.

(c) Consider the Taylor expansion of $g(x)$ around the fixed point $\xi = 4$,

$$g(x) = \xi + \frac{g'(\xi)}{1!}(x-\xi) + \frac{g''(\xi)}{2!}(x-\xi)^2 + \frac{g'''(\eta)}{3!}(x-\xi)^3,$$

where $\eta = \eta(x) \in (\xi, x)$. Since $g'(\xi) = g''(\xi) = 0$, and x_5 is close to ξ:

$$x_6 = g(x_5) \approx 4 + \frac{g'''(\eta)}{3!}(x_5 - 4)^3.$$

Since $g'''(\xi) = -3$, it is reasonable to guess that

$$x_6 \approx 4 - \frac{3}{3!} \cdot (-10^{-6})^3 = 4 + 5 \cdot 10^{-19}.$$

Thus, (iv) is true.

3. (a) We have $f'(x) = 3x^2 - 1$. Hence f' vanishes only at $x_1 = -\frac{1}{\sqrt{3}}$ and $x_2 = \frac{1}{\sqrt{3}}$. Since neither of these points is a zero of f, each of those zeros ξ_i has a neighborhood U_i such that, if $x_0 \in U_i$, then the resulting sequence converges at least quadratically to ξ_i.

Since $f''(x) = 6x > 0$ on $(0, \infty)$, the function is both increasing and convex throughout $(1/\sqrt{3}, \infty)$. Hence, for every $\varepsilon > 0$, if b is sufficiently large, then f satisfies on the interval $[1/\sqrt{3} + \varepsilon, b]$ the sufficient condition ensuring that Newton's method converges to the zero of f when starting at any point in the interval. It follows that $U_3 \supseteq (1, \infty)$.

Thus, (i) is true.

(b) Let $f(x) = e^x - x^2 - 2x$ for $x \in [2.2, 2.3]$. We have

$$f'(x) = e^x - 2x - 2, \qquad x \in [2.2, 2.3],$$

and in particular:

$$f'(\xi) = e^{\xi} - 2\xi - 2 = \xi^2 + 2\xi - 2\xi - 2 = \xi^2 - 2 > 0.$$

Hence Newton's method converges quadratically to ξ when started sufficiently close to it.

However, the situation is different for g_1 and g_2. We have $g_1'(x) = \frac{e^x - 2x}{2}$, so that

$$g_1'(\xi) = \frac{\xi^2 + 2\xi - 2\xi}{2} = \frac{\xi^2}{2} > 1.$$

Hence, starting from a point sufficiently close to ξ, we move farther away from it.

For g_2 we have $g_2'(x) = \frac{2x+2}{x^2+2x}$, so that

$$g_2'(\xi) = \frac{2}{\xi} \cdot \frac{\xi + 1}{\xi + 2} \in (0, 1).$$

Therefore, solving the equation by iterating g_2, starting from a point sufficiently close to ξ, we obtain a sequence converging linearly to ξ.

Thus, (ii) is true.

(c) Obviously, f has a unique root $\xi = 0$. The iteration function corresponding to Newton's method is given by:

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{2}{5}\operatorname{tg} x, \qquad x \in [-\frac{\pi}{4}, \frac{\pi}{4}].$$

Therefore

$$g'(x) = 1 - \frac{2}{5} \cdot \frac{1}{\cos^2 x}, \qquad x \in [-\frac{\pi}{4}, \frac{\pi}{4}],$$

which yields $g'(\xi) = 1 - \frac{2}{5} \cdot \frac{1}{\cos^2 0} = \frac{3}{5}$. Hence, starting from a point near 0 (actually, in our case every point in $[-\pi/4, \pi/4]$ will do), we have linear convergence with $|e_{n+1}| \approx \frac{3}{5}|e_n|$ as $n \to \infty$. This convergence is a bit slower than that of the bisection method, where $|e_{n+1}| \approx \frac{1}{2}|e_n|$.

Thus, (i) is true.