Final #2

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 x be a floating point number belonging to $[2^{-25}, 2^{-22})$.

   (i) If $x \leq 2^{-24}$ then $x \oplus x \odot x = x$, while otherwise $x \oplus x \odot x > x$.
   (ii) If $x \leq 2^{-47/2}$ then $x \oplus x \odot x = x$, while otherwise $x \oplus x \odot x > x$.
   (iii) If $x \leq 2^{-23}$ then $x \oplus x \odot x = x$, while otherwise $x \oplus x \odot x > x$.
   (iv) If $x \leq 2^{-45/2}$ then $x \oplus x \odot x = x$, while otherwise $x \oplus x \odot x > x$.
   (v) None of the above.

   (b) Consider the Matlab code section

```
n=0;
left=1;
right=2;
middle=(left+right)/2;
while((left<middle) && (middle<right))
   if (middle*middle<2)
      left=middle;
   else
      right=middle;
   end;
   middle=(left+right)/2;
   n=n+1;
end
```

We run the code on a system with the specifications listed at the beginning of the question. After the code is run we have

(i) $n = 23, \texttt{middle} \approx 1.33$.

(ii) $n = 23, \texttt{middle} \approx 1.41$.

(iii) $n = 24, \texttt{middle} \approx 1.33$.

(iv) $n = 24, \texttt{middle} \approx 1.41$.

(v) None of the above.

2. (a) Consider the equation:

$$-\arccos\left(e^{\sin x}\right) = x.$$

We use a fixed point iteration with $g(x) = -\arccos\left(e^{\sin x}\right)$ to solve the equation near the solution $\xi = 0$, starting at a point x_0. (Note that the expression on the left-hand side of the equation is defined in a left neighborhood of ξ only. Thus, we shall relate only to such neighborhoods below.)

(i) There is no left neighborhood of ξ with the property that, if x_0 belongs to this neighborhood, then we obtain convergence to ξ.

(ii) There is a left neighborhood of ξ such that, if x_0 belongs to this neighborhood, then we obtain linear convergence to ξ. The convergence is slightly slower than that achieved by the bisection method (in general, when the function is defined in a two-sided neighborhood of the root).

(iii) There is a left neighborhood of ξ such that, if x_0 belongs to this neighborhood, then we obtain linear convergence to ξ. The convergence is slightly faster than that achieved by the bisection method (in general, when the function is defined in a two-sided neighborhood of the root).

(iv) There is a left neighborhood of ξ such that, if x_0 belongs to this neighborhood, then we obtain at least quadratic convergence to ξ.

(v) None of the above.

(b) Consider the equations

$$\ln \cos x - \sin x = 0$$

and

$$\ln \cos x - \sin^2 x = 0.$$

Newton's method is employed to solve the equations. If we start sufficiently close to the common root $\xi = 0$ of the equations then:

(i) The convergence is linear for both equations. For the first equation it is slightly faster than for the bisection method, while for the second it is at almost the same rate.

(ii) The convergence is linear for the first equation and at least quadratic for the second.

(iii) The convergence is at least quadratic for the first equation and linear for the second.

(iv) The convergence is at least quadratic for both equations.

(v) None of the above.

3. (a) Let D be the space of all real continuous functions on the interval $[a, b]$. Denote by D_+ the subset of D, consisting of all those functions having the property that the error in approximating $\int_a^b f(x)dx$, when subdividing the interval in any way and approximating the integral on any of the sub-intervals by the rectangle rule, is positive.

(i) D_+ is closed under multiplication by positive constants and under addition, but not under subtraction and under multiplication.

(ii) D_+ is closed under multiplication by positive constants, but not under addition, but not under subtraction and under multiplication.

(iii) D_+ is closed under addition and subtraction, but not under multiplication.

(iv) D_+ is closed under addition and multiplication, but not under subtraction.

(v) None of the above.

(b) We estimate $\int_0^1 \operatorname{arctg} x dx$ by dividing the interval $[0,1]$ into n sub-intervals of equal length, and using the trapezoid rule for each of them. The error E is approximately

(i) $-\frac{1}{12n^2}$.
(ii) $-\frac{1}{24n^2}$.
(iii) $\frac{1}{24n^2}$.
(iv) $\frac{1}{12n^2}$.
(v) None of the above.

(c) We estimate $\int_1^{2n+1} \ln x dx$ by dividing the interval $[1, 2n+1]$ into n sub-intervals of equal length, and using the midpoint rule for each of them. The total error E satisfies:

(i)
$$-\frac{1}{3}\left(\frac{1}{1^2}+\frac{1}{3^2}+\ldots+\frac{1}{(2n-1)^2}\right) < E < -\frac{1}{3}\left(\frac{1}{3^2}+\frac{1}{5^2}+\ldots+\frac{1}{(2n+1)^2}\right).$$

(ii)
$$-\frac{1}{3}\left(\frac{1}{3^2}+\frac{1}{5^2}+\ldots+\frac{1}{(2n+1)^2}\right) < E < 0.$$

(iii)
$$0 < E < \frac{1}{3}\left(\frac{1}{3^2}+\frac{1}{5^2}+\ldots+\frac{1}{(2n+1)^2}\right).$$

(iv)
$$\frac{1}{3}\left(\frac{1}{3^2}+\frac{1}{5^2}+\ldots+\frac{1}{(2n+1)^2}\right) < E < \frac{1}{3}\left(\frac{1}{1^2}+\frac{1}{3^2}+\ldots+\frac{1}{(2n-1)^2}\right).$$

(v) None of the above.

4. (a) An approximation formula of the form
$$\int_0^1 f(x)dx \approx w_1 f(0) + \frac{1}{4} f(x_2) + w_3 f\left(\frac{5}{8}\right),$$
with some appropriate weights w_1, w_3 and point $x_2 \in [0,1]$, is given, that is completely accurate in case f is a polynomial of degree not exceeding 2. The point x_2 lies in the interval

(i) $[0.1, 0.25]$.

(ii) $(0.25, 0.4]$.

(iii) $(0.4, 0.55]$.

(iv) $(0.55, 0.7]$.

(v) None of the above.

(b) We are looking for an approximation formula of the form

$$\int_{-\pi}^{\pi} f(x)dx \approx w_1 f(x_1) + w_2 f(x_2),$$

with some appropriate weights w_1, w_2 and points $x_1, x_2 \in [-\pi, \pi]$, that will be completely accurate in case f is any function of the form

$$f(x) = a \sin x + b \sin^2 x + c \cos x + d \cos^2 x$$

with constants a, b, c, d. Up to interchanging (w_1, x_1) and (w_2, x_2),

(i) there exists exactly one choice of weights and points.

(ii) there exist exactly two choices of weights and points.

(iii) there exist exactly four choices of weights and points.

(iv) there exist infinitely many choices of weights and points.

(v) None of the above.

Solutions

1. (a) Take $x = 3 \cdot 2^{-26} \in [2^{-25}, 2^{-24}]$. Then

$$\begin{aligned} x \oplus x \odot x &= 2^{-25} \cdot 1.1_2 \oplus 2^{-49} \cdot 1.001_2 \\ &= 2^{-25} \cdot \left(1.1_2 \oplus 2^{-24} \cdot 1.001_2\right) \\ &= 2^{-25} \cdot \left(1.1_2 \oplus 0.\underbrace{00\ldots0}_{23}|1001_2\right) \\ &= 2^{-25} \cdot 1.\underbrace{10\ldots1_2}_{23} \\ &= 3 \cdot 2^{-26} + 2^{48} > x. \end{aligned}$$

   Thus, (v) is true.

   (b) Before the first iteration, the values (in binary) are:

$$\texttt{left} = 1, \texttt{right} = 10, \texttt{middle} = 1.1.$$

   Consider the values at the end of the first few iterations. In order to realize how the numbers behave, we use, in each row, the maximal number of digits required to represent all three numbers, `left`, `right`, `middle`.

n	left	right	middle
1	1.00	1.10	1.01
2	1.010	1.100	1.011
3	1.0110	1.1000	1.0111
4	1.01100	1.01110	1.01101
5	1.011010	1.011100	1.011011

   At each iteration we divide the current interval into two, continuing to work with the one containing $\sqrt{2}$, so obviously $\texttt{middle} \approx \sqrt{2}$. Notice that at the n-th iteration, we need precisely $n+1$ binary digits after the point to represent `middle`, since its last digit is '1'. Furthermore, due to the fact that the first n fractional

digits are identical either to those of `left` or to those of `right`, at the end of the 23-rd round, the program does not enter the 24-th iteration, since `middle` will either be equal to `left` or to `right` at that moment.

Thus, (ii) is true.

2. (a) Since we care only about a left-neighborhood of ξ, to fix ideas we will consider the interval $[-\pi/2, 0]$. We have

$$g'(x) = \frac{e^{\sin x}\cos x}{\sqrt{1 - e^{2\sin x}}}, \qquad x \in [-\pi/2, 0).$$

Obviously,

$$\lim_{x \to \xi^-} g'(x) = \infty,$$

and therefore the fixed point ξ is not attracting.

Thus, (i) is true.

(b) Set

$$f_1(x) = \ln\cos x - \sin x, \qquad f_2(x) = \ln\cos x - \sin^2 x.$$

We have

$$f_1'(x) = -\operatorname{tg} x - \cos x.$$

Substituting $x = \xi$ we obtain $f_1'(\xi) = -1 \neq 0$, and therefore the convergence is at least quadratic for the first equation.

Now we consider the second equation. We have

$$f_2'(x) = -\operatorname{tg} x - \sin 2x$$

and

$$f_2''(x) = -\frac{1}{\cos^2 x} - 2\cos 2x,$$

and in particular $f_2'(\xi) = 0$ and $f_2''(\xi) = -3$. Thus, $\xi = 0$ is a root of f_2 of order 2. The iteration function corresponding to Newton's method is:

$$g_2(x) = x - \frac{f_2(x)}{f_2'(x)}.$$

Now

$$\lim_{x\to\xi} g_2'(x) = \lim_{x\to\xi} \frac{f_2(x)\cdot f_2''(x)}{(f_2'(x))^2},$$

and by l'Hôpital's rule we obtain

$$\begin{aligned}
\lim_{x\to\xi} g_2'(x) &= \lim_{x\to\xi} \frac{f_2'(x)f_2''(x) + f_2(x)f_2^{(3)}(x)}{2f_2'(x)f_2''(x)} \\
&= \frac{1}{2} + \frac{1}{2}\lim_{x\to\xi} \frac{f_2(x)f_2^{(3)}(x)}{f_2'(x)f_2''(x)} \\
&= \frac{1}{2} + \frac{1}{2}\lim_{x\to\xi} \frac{f_2(x)f_2^{(4)}(x) + f_2'(x)f_2^{(3)}(x)}{(f_2''(x))^2 + f_2'(x)f_2^{(3)}(x)} = \frac{1}{2}.
\end{aligned}$$

Hence the convergence is linear for the second equation, with speed almost the same as that of the bisection method.

Thus, (iii) is true.

3. (a) For arbitrary functions $f_1, f_2 \in D_+$, and arbitrary constant $c \in \mathbf{R}$ consider:

$$f(x) = f_1(x) + f_2(x), \qquad x \in [a,b], \tag{1}$$

$$h(x) = cf_1(x), \qquad x \in [a,b]. \tag{2}$$

Obviously, f and h are continuous on the interval $[a,b]$ and the total errors E_f and E_h in approximating the integrals $\int_a^b f(x)dx$ and $\int_a^b h(x)dx$, when subdividing the interval in any way to n sub-intervals and approximating each integral on each of the sub-intervals by the rectangle rule, respectively, are:

$$E_f = \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} f(x)dx - f(x_{i-1})(x_i - x_{i-1}) \right) = \sum_{i=1}^{n} E_{f,i}, \tag{3}$$

and

$$E_h = \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} h(x)dx - h(x_{i-1})(x_i - x_{i-1}) \right) = \sum_{i=1}^{n} E_{h,i}, \tag{4}$$

where $x_0 = a < x_1 < \ldots < x_n = b$ are the division points. By (1) for each subinterval $[x_{i-1}, x_i]$ we have

$$\begin{aligned} E_{f,i} &= \int_{x_{i-1}}^{x_i} (f_1(x) + f_2(x))dx - (f_1(x_{i-1}) + f_2(x_{i-1}))(x_i - x_{i-1}) \\ &= \int_{x_{i-1}}^{x_i} f_1(x)dx - f_1(x_{i-1})(x_i - x_{i-1}) \qquad (5) \\ &\qquad + \int_{x_{i-1}}^{x_i} f_2(x)dx - f_2(x_{i-1})(x_i - x_{i-1}) \\ &= E_{f_1,i} + E_{f_2,i} > 0, \end{aligned}$$

since $f_1, f_2 \in D_+$. Similarly, by (2) we have

$$E_{h,i} = cE_{f_1,i}, \qquad (6)$$

which is positive if and only if $c > 0$. Therefore by (3) and (4) we obtain that D_+ is closed under addition and under multiplication by positive constants. (We mention in passing that (5) and (6) imply, that, for arbitrary fixed division points, the error is a functional linear on the space of continuous functions.)

Consider the following examples to show that D_+ is not closed under subtraction and under multiplication. Put $[a, b] = [0, 1]$ and

$$f_1(x) = f_2(x) = x - 1, \qquad x \in [0, 1].$$

Obviously, $f_1, f_2 \in D_+$. For $x \in [0, 1]$ set

$$f_3(x) = f_1(x) - f_2(x) \equiv 0, \qquad f_4(x) = f_1(x) \cdot f_2(x) = (x - 1)^2,$$

and denote by E_3 and E_4 the total errors corresponding to the approximations of the integrals $\int_0^1 f_3(x)dx$ and $\int_0^1 f_4(x)dx$ by the rectangle rule for any subdivision of $[0, 1]$, respectively. Clearly, $E_3 = 0$, and since $f_4(x)$ is monotonically decreasing on the interval $[0, 1]$, we have $E_4 < 0$. Therefore $f_1 - f_2, f_1 \cdot f_2 \notin D_+$.

Thus, (i) is true.

(b) Let $f(x) = \operatorname{arctg} x$ for $x \in [0, 1]$. We have $f'(x) = \frac{1}{1+x^2}$ and $f''(x) = -\frac{2x}{(1+x^2)^2}$. The total approximation error E of the integral

$\int_0^1 f(x)dx$ by dividing the interval $[0,1]$ into n sub-intervals of equal length, and using the trapezoid rule for each of them is:

$$\begin{aligned} E &= -\sum_{i=1}^{n} f''(\eta_i) \cdot \frac{1}{12n^3}, \qquad (\eta_i \in ((i-1)/n, i/n)) \\ &= \sum_{i=1}^{n} \frac{2\eta_i}{(1+\eta_i^2)^2} \cdot \frac{1}{12n^3} \\ &= \frac{1}{6n^2} \sum_{i=1}^{n} \frac{\eta_i}{(1+\eta_i^2)^2} \cdot \frac{1}{n}. \end{aligned}$$

The sum on the right-hand side is a Riemann sum of the function $g(x) = \frac{x}{(1+x^2)^2}$ on the interval $[0,1]$, multiplied by $\frac{1}{6n^2}$. Thus,

$$E \approx \frac{1}{6n^2} \int_0^1 \frac{x dx}{(1+x^2)^2} = \frac{1}{12n^2} \int_1^2 \frac{dt}{t^2} = \frac{1}{24n^2}.$$

Thus, (iii) is true.

(c) Let $f(x) = \ln x$ for $x \in [1, 2n+1]$. The division points on the interval $[1, 2n+1]$, corresponding to a division of the interval into n sub-intervals of equal length ($h = 2$), are

$$x_0 = 1 < \ldots < x_{i-1} = 2i - 1 < x_i = 2i + 1 < \ldots < x_n = 2n + 1.$$

Thus, the total error is:

$$E = \sum_{i=1}^{n} \frac{f''(\eta_i)}{24} h^3 = \frac{1}{3} \sum_{i=1}^{n} f''(\eta_i), \qquad \eta_i \in (2i-1, 2i+1). \quad (7)$$

We have $f''(x) = -\frac{1}{x^2}$. Since f'' is monotonically increasing on each sub-interval $[2i-1, 2i+1]$, $1 \le i \le n$, we have:

$$-\frac{1}{(2i-1)^2} < f''(\eta_i) < -\frac{1}{(2i+1)^2}, \qquad \eta_i \in (2i-1, 2i+1).$$

Therefore, by (7) we obtain:

$$-\frac{1}{3}\left(\frac{1}{1^2} + \ldots + \frac{1}{(2n-1)^2}\right) < E < -\frac{1}{3}\left(\frac{1}{3^2} + \ldots + \frac{1}{(2n+1)^2}\right).$$

Thus, (i) is true.

4. (a) For the formula in question to be exact for all polynomials up to degree 2, it needs to hold for the polynomials $1,\ x,\ x^2$. Namely, the following equalities need to hold:

$$\begin{cases} w_1 \cdot 1 + \frac{1}{4} \cdot 1 + w_3 \cdot 1 & = & 1, \\ w_1 \cdot 0 + \frac{1}{4} \cdot x_2 + w_3 \cdot \frac{5}{8} & = & \frac{1}{2}, \\ w_1 \cdot 0^2 + \frac{1}{4} \cdot x_2^2 + w_3 \cdot \left(\frac{5}{8}\right)^2 & = & \frac{1}{3}. \end{cases}$$

Equivalently:

$$\begin{cases} 4w_1 + 4w_3 & = & 3, \\ 2x_2 + 5w_3 & = & 4, \\ 48x_2^2 + 75w_3 & = & 64. \end{cases}$$

A routine calculation yields two solutions of the system,

$(a)\ \ w_1 = \frac{9+\sqrt{417}}{120} \approx 0.25, w_3 = \frac{81-\sqrt{417}}{120} \approx 0.50, x_2 = \frac{15+\sqrt{417}}{48} \approx\ 0.74,$ and

$(b)\ \ w_1 = \frac{9-\sqrt{417}}{120} \approx -0.1, w_3 = \frac{81+\sqrt{417}}{120} \approx 0.85, x_2 = \frac{15-\sqrt{417}}{48} \approx\ -0.11.$

Since x_2 should belong to $[0, 1]$, the only feasible solution for our system is the one provided in (a) (with $x_2 \approx 0.74$).
Thus, (v) is true.

(b) For the formula in question to be exact for any function of the form $f(x) = a \sin x + b \sin^2 x + c \cos x + d \cos^2 x$, and using the equality $\cos^2 x = 1 - \sin^2 x$, it needs to hold for the functions $\sin x,\ \cos x,\ \sin^2 x,\ 1$. Namely, the following equalities need to hold:

$$\begin{cases} w_1 \sin x_1 + w_2 \sin x_2 & = & \int_{-\pi}^{\pi} \sin x dx, \\ w_1 \cos x_1 + w_2 \cos x_2 & = & \int_{-\pi}^{\pi} \cos x dx, \\ w_1 \sin^2 x_1 + w_2 \sin^2 x_2 & = & \int_{-\pi}^{\pi} \sin^2 x dx, \\ w_1 \cdot 1 + w_2 \cdot 1 & = & \int_{-\pi}^{\pi} 1 dx. \end{cases}$$

Equivalently:

$$\begin{cases} w_1 \sin x_1 + w_2 \sin x_2 &= 0, \\ w_1 \cos x_1 + w_2 \cos x_2 &= 0, \\ w_1 \sin^2 x_1 + w_2 \sin^2 x_2 &= \pi, \\ w_1 \cdot 1 + w_2 \cdot 1 &= 2\pi. \end{cases}$$

One can easily verify that $x_1, x_2 \neq \pm\pi/2$ and $w_1, w_2 \neq 0$. Thus, moving second term on the left-hand side of each of the first two equations to the right-hand side, and dividing by sides we obtain

$$\operatorname{tg} x_1 = \operatorname{tg} x_2, \qquad x \in [-\pi, \pi].$$

This yields that we have either

$$\text{(I)} \;\; x_1 = \alpha, \;\; x_2 = -\pi + \alpha,$$

or

$$\text{(II)} \;\; x_1 = -\alpha, \;\; x_2 = \pi - \alpha,$$

for some $0 < \alpha < \pi/2$.
Let us start with case (I). By the identities $\sin\alpha = -\sin(-\pi+\alpha)$ and $\cos\alpha = -\cos(-\pi+\alpha)$, the first two equations of the system imply $w_1 = w_2$, and by the last equation we have $w_1 = w_2 = \pi$. Finally, the third equation yields $\sin\alpha = \frac{1}{\sqrt{2}}$, which means $\alpha = \frac{\pi}{4}$. Therefore

$$x_1 = \frac{\pi}{4}, \; x_2 = -\frac{3\pi}{4}, \; w_1 = w_2 = \pi,$$

is the unique solution (up to interchanging (w_1, x_1) and (w_2, x_2)) of the system under case (I).
Similarly one can obtain that

$$x_1 = -\frac{\pi}{4}, \; x_2 = \frac{3\pi}{4}, \; w_1 = w_2 = \pi,$$

is the unique solution (up to interchanging (w_1, x_1) and (w_2, x_2)) of the system under case (II). Therefore, there exist exactly two choices of weights and points.
Thus, (ii) is true.