Final #1

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.

   (a) The largest positive integer n for which $2^n \oplus n$ is a floating point number and $2^n \oplus n > 2^n$ is

   (i) 23.
   (ii) 24.
   (iii) 27.
   (iv) 28.
   (v) None of the above.

   (b) Consider the approximation formula

   $$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

   (h a non-zero number close to 0) for estimating $f'(x)$. Suppose we use the formula to estimate $f'(1)$ for the function $f(x) = \sqrt{x}$. For $k = 1, 2, \ldots$, denote by e_k the absolute value of the error when we take $h = k\varepsilon$. Assume that, when the system is asked to compute $\sqrt{a}$ for some floating point number a, it returns the floating point closest to $\sqrt{a}$.

   (i) $e_1 < e_2 < e_3$.
   (ii) $e_1 > e_2 > e_3$.
   (iii) $e_1 > e_3 > e_2$.

(iv) $e_3 > e_1 > e_2$.

(v) None of the above.

2. (a) Consider the equation:

$$\frac{\pi}{3}\sin x = x.$$

Notice that $\xi = \pi/6$ is a solution and that, for each $\alpha \neq 0$, the equation is equivalent to:

$$\frac{\pi}{3\alpha}\sin x + \frac{\alpha - 1}{\alpha}x = x.$$

Thus, defining

$$g_\alpha(x) = \frac{\pi}{3\alpha}\sin x + \frac{\alpha - 1}{\alpha}x,$$

the original equation may be tackled using a fixed point iteration for any g. Suppose we start from a point sufficiently close to ξ.

(i) If $\alpha > 1 - \frac{\pi}{2\sqrt{3}}$, then the convergence is linear, but becomes slower as α increases. For $\alpha = 1 - \frac{\pi}{2\sqrt{3}}$ the convergence is quadratic. For $\frac{1}{2} - \frac{\pi}{4\sqrt{3}} < \alpha < 1 - \frac{\pi}{2\sqrt{3}}$ the convergence is linear. For $\alpha < \frac{1}{2} - \frac{\pi}{4\sqrt{3}}$ the point ξ is not attracting.

(ii) The convergence is at least linear for every $\alpha \neq 0$ and quadratic for at least one α.

(iii) The convergence is quadratic (or faster) for no $\alpha \neq 0$.

(iv) If $\alpha > \frac{1}{2} - \frac{\pi}{4\sqrt{3}}$, then the convergence is linear. If $\alpha = \frac{1}{2} - \frac{\pi}{4\sqrt{3}}$, then the convergence is at least quadratic.

(v) None of the above.

(b) Newton's method is employed to solve the equation $\cos(\pi e^x) + 1 = 0$. If we start sufficiently close to the root $\xi = 0$ of the equation then:

(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.

3. (a) We approximate $\int_0^{\pi/12} \operatorname{tg} 4x dx$ by dividing the interval $[0, \pi/12]$ into n sub-intervals, not necessarily of equal length, and using one of the rules for each of these intervals. Let E_1 be the total error if the rule used is the rectangle rule, E_2 – if it is the midpoint rule, and E_3 – if it is Simpson's rule. The signs of the errors are as follows.

   (i) $E_1 > 0, E_2 > 0, E_3 > 0$.

   (ii) $E_1 < 0, E_2 > 0, E_3 > 0$.

   (iii) $E_1 > 0, E_2 > 0, E_3 < 0$.

   (iv) The sign of at least one of the E_i's depends in a non-trivial way on n and the division points.

   (v) None of the above.

   (b) We estimate $\int_0^1 \ln(x(x+1))dx$ by dividing the interval $[0, 1]$ into n sub-intervals of equal length, and using the rectangle rule for each of them, but with the right endpoint of each sub-interval instead of its left endpoint. Let E be the error. For sufficiently large n

   (i) $|E|$ becomes arbitrarily large.

   (ii) $|E| \approx \frac{C}{n}$ for some constant $C > 0$.

   (iii) $|E| \approx \frac{\ln 4\pi n}{2n}$.

   (iv) $|E| \approx \frac{C}{\ln n}$ for some constant $C > 0$.

   (v) None of the above.

   (c) We estimate $\int_0^{\pi/3} \sqrt{\cos x} dx$ by dividing the interval $[0, \pi/3]$ into n sub-intervals of equal length, and using the midpoint rule for each of them. Let E be the error. Then:

   (i) $E \approx -\frac{\pi^2\sqrt{6}}{864n^2}$.

   (ii) $E \approx -\frac{\pi^2\sqrt{2}}{864n^2}$.

   (iii) $E \approx \frac{\pi^2\sqrt{2}}{864n^2}$.

(iv) $E \approx \frac{\pi^2\sqrt{6}}{864n^2}$.
(v) None of the above.

4. We are interested in finding an approximation formula of the form

$$\int_0^1 f(x)dx \approx w_1 f(1/3) + w_2 f(x_2),$$

with some appropriate weights w_1, w_2 and point $x_2 \in [0,1]$, that will be completely accurate in case f is a polynomial of degree not exceeding 2.

(a) We must choose:
(i) $w_1 = w_2 = 1/2, x_2 = 2/3$.
(ii) $w_1 = 1/3, w_2 = 2/3, x_2 = 1/2$.
(iii) $w_1 = w_2 = 1/2, x_2 = 1/2$.
(iv) $w_1 = 3/4, w_2 = 1/4, x_2 = 1$.
(v) None of the above.

(b) Suppose there exist w_1, w_2, x_2 for which the above requirements are satisfied. Let $\langle \cdot, \cdot \rangle$ be the inner product defined on the space of all real polynomials by

$$\langle Q_1, Q_2 \rangle = \int_0^1 Q_1(x)Q_2(x)dx, \qquad Q_1, Q_2 \in \mathbf{R}[x].$$

Consider the polynomials

$$P_1(x) = x - x_2, \qquad P_2(x) = (x - 1/3)(x - x_2).$$

(i) Neither one of the polynomials P_i is orthogonal to all constant polynomials.
(ii) The polynomial P_1 is not orthogonal to all constant polynomials. The polynomial P_2 is orthogonal to all constant polynomials, but not to all polynomials of degree not exceeding 1.
(iii) The polynomial P_1 is orthogonal to all polynomials of degree not exceeding 1, but not to all polynomials of degree not exceeding 2. The polynomial P_2 is orthogonal to all constant polynomials, but not to all polynomials of degree not exceeding 1.

(iv) The polynomial P_1 is orthogonal to all polynomials of degree not exceeding 2, but not to all polynomials of degree not exceeding 3. The polynomial P_2 is orthogonal to all polynomials of degree not exceeding 1, but not to all polynomials of degree not exceeding 2.

(v) None of the above.

Solutions

1. (a) Integers in the range [16, 31] are of the form $1.\underbrace{b_1b_2b_3b_40\ldots0}_{23}\cdot 2^4$.

For any $n \in [16,\ 31]$, the addition of n to 2^n requires shifting the representation of n by $n-4$ bits to the right, to obtain both numbers represented with the same exponent. As long as $n - 4 \leq 23$, we clearly have $2^n \oplus n > 2^n$, because the most significant digit (the implicit 1 to the left of the binary point) of n is still shifted to one of the first 23 digits after the binary point. For $n = 28$, which requires 24 shifts, the representation is $0.\underbrace{0\ldots0}_{23}111 \cdot 2^{28}$. Hence

$$2^{28} + 28 = 1 \cdot 2^{28} + 0.\underbrace{0\ldots0}_{23}111 \cdot 2^{28} = 1.\underbrace{0\ldots0}_{23}111 \cdot 2^{28},$$

which is rounded to

$$1.\underbrace{0\ldots0}_{22}1 \cdot 2^{28} = 2^{28} + 2^5 > 2^{28}.$$

For $n > 28$, the shift will be of at least 25 places to the right to obtain the same exponent for n, leading to $0.\underbrace{0\ldots0}_{23}01b_1\ldots \cdot 2^n$ (with the leading 1 where shown or even farther to the right). It follows that $2^n \oplus n = 2^n$.

Thus, (iv) is true.

(b) Clearly, $f'(1) = \frac{1}{2}$. Using Taylor's approximation for $f(x+h)$, where $x = 1$, we obtain $f(1+h) = \sqrt{1+h} \approx 1 + \frac{1}{2}(1+h-1)$. For single precision, $\varepsilon = 2^{-23}$, so h assumes the values $2^{-23}, 2 \cdot 2^{-23}, 3 \cdot 2^{-23}$, which will be used for evaluating e_1, e_2, e_3, respectively. The general expression we are interested in is

$$e_k = \left|f'(1) - \text{round}\left(\left(f(1 \oplus k \otimes \varepsilon) \ominus f(1)\right) \oslash (k \otimes \varepsilon)\right)\right|, \quad k = 1, 2, 3.$$

(In fact, when we write $f(a)$ for some floating point number a, we refer to the approximation provided for $f(a)$ by the system.) Hence:

$$e_k \approx \left|\frac{1}{2} - \text{round}\left(\left(1 \oplus \frac{1}{2} \otimes (1 \oplus k \otimes 2^{-23} \ominus 1) \ominus 1\right) \oslash \left(k \otimes 2^{-23}\right)\right)\right|,$$

for $k = 1, 2, 3$. Now we complete the calculation for each k separately:

- $k = 1$:

$$\begin{aligned}
&\text{round}\left(\left(1 \oplus \frac{1}{2} \otimes (1 \oplus 2^{-23} \ominus 1) \ominus 1\right) \oslash 2^{-23}\right) \\
&= \text{round}\left(\left(1 \oplus \frac{1}{2} \otimes 2^{-23} \ominus 1\right) \oslash 2^{-23}\right) \\
&= \text{round}\left(\left(1 \oplus 2^{-24} \ominus 1\right) \oslash 2^{-23}\right) \\
&= \text{round}\left(0 \oslash 2^{-23}\right) = 0.
\end{aligned}$$

  Thus,

$$e_1 = \left|\frac{1}{2} - 0\right| = \frac{1}{2}.$$

- $k = 2$:

$$\begin{aligned}
&\text{round}\left(\left(1 \oplus \frac{1}{2} \otimes (1 \oplus 2 \otimes 2^{-23} \ominus 1) \ominus 1\right) \oslash \left(2 \otimes 2^{-23}\right)\right) \\
&= \text{round}\left(\left(1 \oplus \frac{1}{2} \otimes 2^{-22} \ominus 1\right) \oslash 2^{-22}\right) \\
&= \text{round}\left(\left(1 \oplus 2^{-23} \ominus 1\right) \oslash 2^{-22}\right) \\
&= \text{round}\left(2^{-23} \oslash 2^{-22}\right) = \frac{1}{2}.
\end{aligned}$$

  Thus,

$$e_2 = \left|\frac{1}{2} - \frac{1}{2}\right| = 0.$$

- $k = 3$:

$$\begin{aligned}
&\text{round}\left(\left(1 \oplus \frac{1}{2} \otimes (1 \oplus 3 \otimes 2^{-23} \ominus 1) \ominus 1\right) \oslash \left(3 \otimes 2^{-23}\right)\right) \\
&= \text{round}\left(\left(1 \oplus \frac{1}{2} \otimes 3 \otimes 2^{-23} \ominus 1\right) \oslash \left(3 \otimes 2^{-23}\right)\right) \\
&= \text{round}\left(\left(1 \oplus 3 \otimes 2^{-24} \ominus 1\right) \oslash \left(3 \otimes 2^{-23}\right)\right) \\
&= \text{round}\left(2^{-22} \oslash \left(3 \otimes 2^{-23}\right)\right) \\
&= \text{round}\left(\frac{2}{3}\right).
\end{aligned}$$

Thus,

$$e_3 = \left|\frac{1}{2} - \text{round}\left(\frac{2}{3}\right)\right| \approx \frac{1}{6}.$$

Thus, (iii) is true.

2. (a) We have

$$g'_\alpha(x) = \frac{\pi}{3\alpha}\cos x + \frac{\alpha - 1}{\alpha}, \qquad (\alpha \neq 0),$$

and substituting $\xi = \pi/6$ we obtain:

$$g'_\alpha(\pi/6) = \frac{\pi}{3\alpha}\cos\frac{\pi}{6} + \frac{\alpha - 1}{\alpha} = 1 - \frac{1}{\alpha}\left(1 - \frac{\pi}{2\sqrt{3}}\right), \qquad (\alpha \neq 0). \tag{1}$$

If $\alpha = 1 - \frac{\pi}{2\sqrt{3}}$ then $g'_\alpha(\pi/6) = 0$, so that the convergence is quadratic. If $\alpha > 1 - \frac{\pi}{2\sqrt{3}}$ then $0 < g'_\alpha(\pi/6) < 1$, so that the convergence is linear. In this case the error decreases (almost) as a geometric series with ratio $q = g'_\alpha(\pi/6)$, and since the right-hand hide of (1) increases with α in this range, therefore the convergence becomes slower as α increases. If $\frac{1}{2} - \frac{\pi}{4\sqrt{3}} < \alpha < 1 - \frac{\pi}{2\sqrt{3}}$ then $-1 < g'_\alpha(\pi/6) < 0$, and the convergence is again linear. If $\alpha < \frac{1}{2} - \frac{\pi}{4\sqrt{3}}$ then $g'_\alpha(\pi/6) < -1$. Since $|g'_\alpha(\pi/6)| > 1$, and the fixed point ξ is not attracting.
Thus, (i) is true.

(b) We have

$$f'(x) = -\pi e^x \sin(\pi e^x)$$

and

$$f''(x) = -\pi e^x \sin(\pi e^x) - (\pi e^x)^2 \cos(\pi e^x),$$

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

$$g(x) = x - \frac{f(x)}{f'(x)} = x + \frac{\cos(\pi e^x) + 1}{\pi e^x \sin(\pi e^x)}.$$

Now

$$g'(x) = \frac{(\pi e^x \cos(\pi e^x) + \sin(\pi e^x))e^{-x}}{(\cos(\pi e^x) - 1)\pi}, \tag{2}$$

and a routine calculation yields $g'(x) = \lim_{x \to 0} g'(x) = \frac{1}{2}$. Hence the convergence is linear, with speed almost the same as that of the bisection method.

Thus, (ii) is true.

3. (a) Let $x_0 = 0 < x_1 < \ldots < x_n = \pi/12$ be the division points. The errors $E_{1,i},\ E_{2,i}$, and $E_{3,i}$ in each sub-interval $[x_{i-1}, x_i],\ \ 1 \leq i \leq n$, when using the rectangle rule, the midpoint rule and Simpson's rule, respectively, are:

$$\begin{aligned}
E_{1,i} &= f'(\eta_{1,i})\frac{(x_i - x_{i-1})^2}{2}, && \eta_{1,i} \in (x_{i-1}, x_i),\\
E_{2,i} &= f''(\eta_{2,i})\frac{(x_i - x_{i-1})^3}{24}, && \eta_{2,i} \in (x_{i-1}, x_i),\\
E_{3,i} &= -f^{(4)}(\eta_{3,i})\frac{(x_i - x_{i-1})^5}{90 \cdot 2^5}, && \eta_{3,i} \in (x_{i-1}, x_i).
\end{aligned}$$

The corresponding total errors are:

$$E_1 = \sum_{i=1}^{n} E_{1,i}, \qquad E_2 = \sum_{i=1}^{n} E_{2,i}, \qquad E_3 = \sum_{i=1}^{n} E_{3,i}.$$

One verifies by induction that $f^{(k)}(x)$ is a polynomial of degree $k+1$ with non-negative coefficients in $\operatorname{tg} 4x$ for each $k \geq 0$. For example,

$$f'(x) = \frac{4}{\cos^2 4x} = 2^2(\operatorname{tg}^2 4x + 1),$$
$$f''(x) = 2^5(\operatorname{tg}^3 4x + \operatorname{tg} 4x),$$

and

$$f^{(4)}(x) = 2^{11}\left(3\operatorname{tg}^5 4x + 5\operatorname{tg}^3 4x + 2\operatorname{tg} 4x\right).$$

In particular, since $\operatorname{tg} 4x$ is positive throughout the interval, so is $f^{(k)}(x)$ for every k. Hence, $E_1 > 0, E_2 > 0, E_3 < 0$.

Thus, (iii) is true.

(b) Since $\ln(x(x+1)) = \ln x + \ln(x+1)$, we have:

$$\int_0^1 \ln(x(x+1))dx = \int_0^1 \ln x dx + \int_0^1 \ln(x+1)dx. \tag{3}$$

Moreover, when approximating the left-hand side of (3) by the rectangle rule (or any other rule for that matter) we obtain the sum of the approximations obtained for the two integrals on the right-hand side. Note that the first integral on the right-hand side of (3) was studied in class. When using the rectangle rule with the right endpoint of each sub-interval instead of its left endpoint, it is approximated as follows:

$$\int_0^1 \ln x dx \approx \frac{1}{n}\sum_{i=1}^{n} \ln\frac{i}{n} = \frac{1}{n}\ln n! - \ln n. \tag{4}$$

Similarly, for the second integral on the right-hand side of (3) we have:

$$\begin{aligned}\int_0^1 \ln(x+1)dx &\approx \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{i}{n}+1\right)\\ &= \frac{1}{n}\ln(2n)! - \frac{1}{n}\ln n! - \ln n. \end{aligned}\tag{5}$$

Substituting (4) and (5) in the right-hand side of (3), we obtain:

$$\int_0^1 \ln(x(x+1))dx \approx \frac{1}{n}\ln(2n)! - 2\ln n. \tag{6}$$

Now, by Stirling's formula $(2n)! \approx \sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n}$, and therefore:

$$\int_0^1 \ln(x(x+1))dx \approx \frac{1}{2n}\ln 4\pi n + 2\ln 2 - 2. \tag{7}$$

Since $\int \ln x dx = x\ln x - x + c$, we have

$$\begin{aligned}\int_0^1 \ln(x(x+1))dx &= [x\ln x - x + (x+1)\ln(x+1) - (x+1)]_0^1 \\ &= 2\ln 2 - 2. \end{aligned} \tag{8}$$

By (7) and (8):

$$E \approx -\frac{1}{2n}\ln 4\pi n.$$

Thus, (iii) is true.

(c) Let $f(x) = \sqrt{\cos x}$. For $1 \le i \le n$, the error in the sub-interval $[\frac{\pi(i-1)}{3n}, \frac{\pi i}{3n}]$ is:

$$E_i = \frac{f''(\eta_i)}{24}\cdot\left(\frac{\pi}{3n}\right)^3, \qquad \left(\eta_i \in \left(\frac{\pi(i-1)}{3n}, \frac{\pi i}{3n}\right)\right).$$

Hence the total error is:

$$E = \sum_{i=1}^n E_i = \sum_{i=1}^n \frac{f''(\eta_i)}{24}\left(\frac{\pi}{3n}\right)^3 = \frac{1}{24}\left(\frac{\pi}{3n}\right)^2 \sum_{i=1}^n f''(\eta_i)\cdot\frac{\pi}{3n}.$$

The sum on the right-hand side is a Riemann sum of the function f'' on the interval $[0, \frac{\pi}{3}]$. Thus,

$$E \approx \frac{1}{24}\left(\frac{\pi}{3n}\right)^2 \int_0^{\pi/3} f''(x)dx = \frac{1}{24}\left(\frac{\pi}{3n}\right)^2\left(f'\left(\frac{\pi}{3}\right) - f'(0)\right). \tag{9}$$

Now $f'(x) = -\frac{\sin x}{2\sqrt{\cos x}}$, so that (9) yields

$$E \approx \frac{1}{24}\left(\frac{\pi}{3n}\right)^2\left(-\frac{\sin \pi/3}{2\sqrt{\cos \pi/3}} + \frac{\sin 0}{2\sqrt{\cos 0}}\right) = -\frac{\pi^2\sqrt{6}}{864n^2}.$$

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:

$$\begin{cases} w_1 \cdot 1 + w_2 \cdot 1 & = & 1 \\ w_1 \cdot \frac{1}{3} + w_2 \cdot x_2 & = & \frac{1}{2} \\ w_1 \cdot \frac{1}{9} + w_2 \cdot x_2^2 & = & \frac{1}{3} \end{cases}$$

A routine calculation shows that the choice $x_2 = 1,\ w_1 = 3/4$ and $w_2 = 1/4$ indeed yields a solution of the system.

Thus, (iv) is true.

(b) P_1 is not orthogonal to all constant polynomials. In fact

$$\begin{aligned} \langle P_1, 1 \rangle &= \int_0^1 P_1(x)dx \\ &= w_1 P_1(1/3) + w_2 P_1(x_2) \\ &= w_1 P_1(1/3) = 3/4 \cdot (1/3 - 1) \neq 0. \end{aligned}$$

The polynomial P_2 is orthogonal to all constant polynomials. Indeed, for any constant $c \in \mathbf{R}$

$$\langle P_2, c \rangle = \int_0^1 cP_2(x)dx = w_1 c P_2(1/3) + w_2 c P_1(x_2) = 0.$$

However, P_2 is not orthogonal to all polynomials of degree not exceeding 1. For example,

$$\langle P_2, x - 1/3 \rangle = \int_0^1 (x - 1/3)^2 (x - 1)dx < 0,$$

since the only zeros of the integrand $(x - 1/3)^2(x - 1)$ are $x = 1/3$ or $x = 1$, and for all other values of $x \in [0, 1]$ it is negative.

Thus, (ii) is true.