Brief Solution to Midterm 01

Oct 27, 2020.

1. (10 pts) The (fictional) one-and-half precision format uses 48 bits to store a binary floating point number of the form $\pm 1.a_1a_2\cdots a_t \times 2^e$ where $a_j \in \{0,1\}$, $-510 \le e \le 511$. Find t and derive an upper bound for relative error caused by rounding. Express your final answer as a real number, but need not convert it to decimal expression.

   Ans:

   There are total 1022 different exponents ($-510 \le e \le 511$).
   It takes 10 bits to give 1022 or more different exponents ($2^{10} = 1024$). **(2 pts)**
   Total bits $= 1 + t + 10 = 48 \Rightarrow t = 37$ **(2 pts)**.

   Let $x = \pm 1.a_1a_2\cdots a_t \ldots \times 2^e$.
   If $a_{t+1} = 0$, then $fl_{round}(x) = \pm 1.a_1a_2\cdots a_t \times 2^e$. A bound for the relative error is

$$\frac{|x - fl_{round}(x)|}{|x|} = \frac{|0.a_{t+1}a_{t+2}\ldots|}{|1.a_1a_2\ldots a_ta_{t+1}\ldots|} \times 2^{-t} \le 2^{-(t+1)}. \textbf{ (2 pts)}$$

   If $a_{t+1} = 1$, then $fl_{round}(x) = \pm(1.a_1a_2\cdots a_t + 2^{-t}) \times 2^e$. The upper bound for relative error becomes

$$\frac{|x - fl_{round}(x)|}{|x|} = \frac{|1 - 0.a_{t+1}a_{t+2}\ldots|}{|1.a_1a_2\ldots a_ta_{t+1}\ldots|} \times 2^{-t} \le 2^{-(t+1)}. \textbf{ (2 pts)}$$

   Therefore, an upper bound for relative error caused by rounding is 2^{-38} **(2 pts)**.

2. (10 pts) Given p_0, p_1 and p_2, the general solution to the recursion formula $p_n = \frac{10}{3}p_{n-1} - 3p_{n-2} + \frac{2}{3}p_{n-3}$ is $p_n = c_1 1^n + c_2 2^n + c_3(\frac{1}{3})^n$ (need not show this). Find all $(c_1, c_2, c_3) \neq (0,0,0)$ such that the above iteration is unstable in relative error. Explain.

   Ans: Relative error $= \left|\frac{e_n}{p_n^{\text{exact}}}\right|$. (**3 pts**)

   Note that

$$e_n \approx= d_1 1^n + d_2 2^n + d_3(\frac{1}{3})^n \qquad \textbf{(3pts)}$$

   d_1, d_2 and d_3 are of $O(p_i - fl(p_i))$, $i = 0,1,2$. Therefore d_1, d_2 and d_3 are of $O(\varepsilon_M)$.**(1 pts)**.

   Therefore, relative error $= \frac{O(\varepsilon_M)1^n + O(\varepsilon_M)2^n + O(\varepsilon_M)(\frac{1}{3})^n}{c_1 1^n + c_2 2^n + c_3(\frac{1}{3})^n}$. It is stable if and only if $c_2 \neq 0$. (**3 pts**)

3. (10 pts) The first few iteration $(p_i, f(p_i))$, $i = 0,1,2,3,4$ of method of false position for some equation $f(x) = 0$ is given by

$$(0,-2), \quad (3,1), \quad (*,2), \quad (*,1), \quad (*,\frac{2}{9})$$

Find p_5 (4 digits will do). Explain.

Ans:

$$f(p_1)f(p_0) < 0 \Rightarrow a = p_0,\ b = p_1,\ p_2 = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)} = 2$$
$$f(p_2)f(p_0) < 0 \Rightarrow a = p_0,\ b = p_2,\ p_3 = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)} = 1$$
$$f(p_3)f(p_0) < 0 \Rightarrow a = p_0,\ b = p_3,\ p_4 = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)} = \frac{2}{3}$$
$$f(p_4)f(p_0) < 0 \Rightarrow a = p_0,\ b = p_4.\ p_5 = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)} = 0.6$$

4. (15 pts) Use any method to find a solution of $\sqrt{1+0.9x} - \sqrt{1-0.8x} = 1.0 \times 10^{-10}$ to 15 correct digits. You need to prevent loss of accuracy. Standard methods only gives you about 5 correct digits (and 1/3 partial credits).

   Ans:

   Apply the following identity

   $$a^2 - b^2 = (a+b)(a-b)$$

   that avoids the subtraction of two nearly identical numbers and gives

   $$f(x) = \frac{1.7x}{\sqrt{1+0.9x} + \sqrt{1-0.8x}} - 10^{-10}. \textbf{ (7 pts)}$$

   Then solve $f(x) = 0$ by any numerical method to find the solution

   $$x_* \approx 1.17647058823875 \times 10^{-10}. \textbf{ (8 pts)}$$

5. (10+5 pts) It is known that the unique solution to $f(x) = x + 3\sin(x) - 0.01 = 0$ is located near $x = 0$.

   (a) Find a fixed point iteration that will converge for any $x_0 \in [-\frac{1}{2}, \frac{1}{2}]$. Show that your method satisfies the assumptions of a relevant Theorem, but need not prove the Theorem again. You can use the numerical values of $\sin(\frac{1}{2})$, $\cos(\frac{1}{2})$, $\exp(\frac{1}{2})$, etc. in your proof.

   (b) Find an N (need not be optimal) such that $|x_n - x^*| < 10^{-30}$ for all $n \geq N$ with $x_0 = 0$ (assuming a higher precision floating point arithmetic is used).

   Ans:

(a) Direct fixed point iteration with $x^{(k+1)} = g_0^{(k)}(x) = 0.01 - 3\sin(x^{(k)})$ does not converge. Instead, a proper choice of α and $g(x) = \alpha x + (1-\alpha)g_0(x)$ will result in local convergence **(2 pts)**. One could choose

$$\alpha = \frac{g_0'(\xi)}{g_0'(\xi) - 1}$$

for some ξ near 0. Since $\xi \approx 0$, $g_0'(\xi) \approx -3$, we take $\alpha = \frac{-3}{-3-1} = \frac{3}{4}$. **(2 pts)**
Since $g(x) = \frac{3}{4}(x - \sin x) + 0.0025$, $g'(x) = \frac{3}{4}(1 - \cos(x))$
Therefore

$$0 < \frac{3}{4}\left(1 - \cos(\frac{1}{2})\right) \le g'(x) \le \frac{3}{4} \quad \text{on } [-\frac{1}{2}, \frac{1}{2}].$$

It follows that g is an increasing function on $[-\frac{1}{2}, \frac{1}{2}]$,

$$-\frac{1}{2} < -0.012931... = g(-\frac{1}{2}) \le g(x) \le g(\frac{1}{2}) = 0.017931... < \frac{1}{2}$$

(thus $g([-\frac{1}{2}, \frac{1}{2}]) \subset [-\frac{1}{2}, \frac{1}{2}]$ **(3 pts)**) and

$$|g'(x)| = \left|\frac{3}{4}(1 - \cos(x))\right| \le \frac{3}{4} = k < 1 \ \forall\ x \in (-\frac{1}{2}, \frac{1}{2}) \ \textbf{(3 pts)}$$

(b) Correct estimate **(3 pts)** and correct N **(2 pts)**.
[Method 1]

$$|x_n - x_*| \le k^n \max\{x_0 - a,\ b - x_0\} = \left(\frac{3}{4}\right)^n \max\left\{0 - \left(-\frac{1}{2}\right),\ \frac{1}{2} - 0\right\} = \left(\frac{3}{4}\right)^n \frac{1}{2} < 10^{-30}$$

$$\Rightarrow n > \frac{\log_{10} 2 - 30}{\log_{10} \frac{3}{4}} = 237.71... \Rightarrow N = 238.$$

[Method 2]

$$|x_n - x_*| \le \frac{k^n}{1-k}|x_1 - x_0| = 4\left(\frac{3}{4}\right)^n |g(0) - 0| = \left(\frac{3}{4}\right)^n 0.01 < 10^{-30}$$

$$\Rightarrow n > \frac{28}{\log_{10} \frac{4}{3}} = 224.11... \Rightarrow N = 225.$$

6. (15 pts) Give a (at least) cubically convergent method to solve for $e^x - 1 = 0$. Give the formula and prove that it is at least cubically convergent (locally). If you cannot do it, do the same for a locally (at least) quadratically convergent method for 1/3 partial credit.

Answer:

[Cubic]

One solution is given by (there may be others)

$$x_{n+1} = g(x_n), \quad g(x) = x - \frac{f(x)}{f'(x)} - \frac{f''(x)}{2f'(x)} \left[\frac{f(x)}{f'(x)}\right]^2. \quad \textbf{(5 pts)}$$

Check that $g'(p) = g''(p) = 0$ and $g'''(p) \neq 0$.

Therefore

$$p_{n+1} - p = g(p_n) - g(p) = \frac{g'''(\xi)}{3!}(p_n - p)^3$$

and

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^3} = \frac{|g^{(3)}(p)|}{3!}. \quad \textbf{(10 pts)}$$

[Quadratic]

$$x_{n+1} = g(x_n), \quad g(x) = x - \frac{f(x)}{f'(x)}. \quad \textbf{(2 pts)}$$

Check that $g'(p) = 0$ and $g''(p) \neq 0$.

Therefore

$$p_{n+1} - p = g(p_n) - g(p) = \frac{g''(\xi)}{2!}(p_n - p)^2$$

and

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2!}. \quad \textbf{(3 pts)}$$

7. (15 pts) One way of computing π is given by the Wallis formula

$$\frac{\pi}{2} = \prod_{n=1}^{\infty} \left(\frac{(2n)^2}{(2n-1)(2n+1)}\right).$$

The N-term approximation is therefore given by

$$\frac{\pi_N}{2} = \left(\frac{2\cdot 2}{1\cdot 3}\right)\left(\frac{4\cdot 4}{3\cdot 5}\right)\left(\frac{6\cdot 6}{5\cdot 7}\right)\cdots\left(\frac{2N\cdot 2N}{(2N-1)\cdot(2N+1)}\right)$$

To prevent overflow too quickly, it is better to evaluate the last multiplication as $*(2N)/(2N-1) * (2N)/(2N+1)$. Find the rate of convergence of $\lim_{n\to\infty} \pi_n = \pi$ numerically. Extra points without using the limit π explicitly.

Hint: The convergence is slow, try not to produce all the data points. For example, π_{100}, π_{200}, $\cdots$, π_{10000} should be enough to analyze.

To check if an integer j is a multiple of 100 or not, you can use mod or check round$(j/100) * 100$.

Ans:

Programming:

Completed program that can produce all π_n needed: (**10 pts**)

Analysis:

Method 1:

Try semilogy and loglog plot of $|\pi_n - \pi|$ vs n to determine whether $|\pi_n - \pi| \approx Cn^{-p}$ or $|\pi_n - \pi| \approx C\alpha^{-n}$ or something else.

The results indicates that $|\pi_n - \pi| \approx Cn^{-p}$ where $-p$ is the slope in the loglog plot and $p \approx 1$. Comparing the loglog plot of $|\pi_n - \pi|$ vs n, with the loglog plot of n^{-1} vs n confirms that $p = 1$ (**5 pts**).

Method 2 (extra 5 pts):

Proceed to find out the constants C, p or C, α.

Directly try $\pi_n - \pi \approx Cn^{-p}$ and find p through

$$p \approx \log_2 \frac{a_N - a_{2N}}{a_{2N} - a_{4N}}$$

Different choices of $N = 300, 500, \cdots, 1000, 1500$ all give consistent answer $p \approx 1$.

Conclusion: $\pi_n - \pi = O(n^{-1})$

8. (15 pts) Use any method to solve the nonlinear system of equations

$$\sin(x) + \frac{2y}{1+x} = 0.01, \qquad 5x + \sin(\frac{6y}{1+y^2}) = 0.02.$$

Write your answer in the format of 'format long e'.

Hint: the solution is near $(0, 0)$ where $\sin x \approx x$, $\frac{y}{1+x} \approx y$, etc. to leading order.

Ans:

Method 1:

Let

$$g_1(\mathbf{x}) = \frac{1}{5}\left(0.02 - \sin\left(\frac{6y}{1+y^2}\right)\right)$$
$$g_2(\mathbf{x}) = \frac{1}{2}(0.01 - \sin(x))(1+x)$$

and

$$\mathbf{G}(\mathbf{x}) = (g_1(\mathbf{x}),\ g_2(\mathbf{x}))^{\mathrm{T}}$$
$$\bar{\mathbf{G}}(\mathbf{x}) = \alpha\mathbf{x} + (I - \alpha)\mathbf{G}(\mathbf{x})$$

where α is a 2×2 matrix and I is the identity matrix. From the Hint, we can take $\mathbf{x}_0 = (0,\ 0)^{\mathrm{T}}$ and

$$\alpha = (D\mathbf{G}(\mathbf{x}_0) - I)^{-1} D\mathbf{G}(\mathbf{x}_0).$$

Method 2:

From the Hint, the linear approximation of the left hand side is $(x + 2y, 5x + 6y)$, so we can rewrite the equation as

$$x + 2y = 0.01 + x + 2y - \sin(x) - \frac{2y}{1+x} \equiv h_1(x, y),$$

$$5x + 6y = 0.02 + 6y - \sin(\frac{6y}{1+y^2}) \equiv h_2(x, y).$$

This suggests the fixed point iteration:

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 5 & 6 \end{pmatrix}^{-1} \begin{pmatrix} h_1(x^{(k)}, y^{(k)}) \\ h_2(x^{(k)}, y^{(k)}) \end{pmatrix}$$

Method 3: Newton's Method.

(Correct algorithm of any convergent method = 10 pts)

If $\mathbf{x}_0 = (0,\ 0)^{\mathrm{T}}$, then the iteration gives

$$\mathbf{x}_* \approx (-4.882452175e-03,\ 7.404885005e-03)^{\mathrm{T}}. \textbf{ (5 pts)}$$