Midterm 01

Oct 24, 2017.

1. (10 pts) The half precision format uses 16 bits to store a binary floating point number of the form $\pm 1.a_1 a_2 \cdots a_t \times 2^e$ where $a_j \in \{0, 1\}$, $-14 \leq e \leq 15$. Find t and <u>derive</u> an upper bound for relative error caused by <u>rounding</u>. Express your final answer as a real number, but need not convert it to decimal expression.

   Ans:

   There are total 30 different exponents ($-14 \leq e \leq 15$).
   It takes 5 bits to give 30 or more different exponents ($2^5 = 32$). (**2 pts**)
   Total bits $= 1 + t + 5 = 16 \Rightarrow t = 10$ (**2 pts**).

   Let $x = \pm 1.a_1 a_2 \cdots a_t \ldots \times 2^e$.
   If $a_{t+1} = 0$, then $fl_{round}(x) = \pm 1.a_1 a_2 \cdots a_t \times 2^e$. A bound for the relative error is

   $$\frac{|x - fl_{round}(x)|}{|x|} = \frac{|0.a_{t+1}a_{t+2}\ldots|}{|1.a_1 a_2 \ldots a_t a_{t+1} \ldots|} \times 2^{-t} \leq 2^{-(t+1)}. \text{ (2 pts)}$$

   If $a_{t+1} = 1$, then $fl_{round}(x) = \pm(1.a_1 a_2 \cdots a_t + 2^{-t}) \times 2^e$. The upper bound for relative error becomes

   $$\frac{|x - fl_{round}(x)|}{|x|} = \frac{|1 - 0.a_{t+1}a_{t+2}\ldots|}{|1.a_1 a_2 \ldots a_t a_{t+1} \ldots|} \times 2^{-t} \leq 2^{-(t+1)}. \text{ (2 pts)}$$

   Therefore, an upper bound for relative error caused by rounding is 2^{-11} (**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. <u>Explain</u>.

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

   Note that

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

   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 $O(\delta)$.(**1 pts**).

   Therefore, relative error $= \frac{O(\delta)1^n + O(\delta)2^n + O(\delta)(\frac{1}{3})^n}{c_1 1^n + c_2 2^n + c_3(\frac{1}{3})^n}$. 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

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

Find p_5 (4 digits will do). <u>Explain.</u>

Ans:

$$f(p_1)f(p_0) < 0 \Rightarrow a = p_0, \ b = p_1$$
$$f(p_2)f(p_0) < 0 \Rightarrow a = p_0, \ b = p_2$$
$$f(p_3)f(p_0) < 0 \Rightarrow a = p_0, \ b = p_3$$
$$f(p_4)f(p_0) < 0 \Rightarrow a = p_0, \ b = p_4$$

(up to here = 4 pts)

$$\Rightarrow p_5 = p_4 - f(p_4)\frac{p_4 - p_0}{f(p_4) - f(p_0)} \ \textbf{(4 pts)} = 0.6 \ \textbf{(2 pts)}.$$

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:

2

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

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

for some ξ near 0. If $\xi = 0$, then $\beta = \frac{3}{4}$. **(2 pts)**

First check $g([-\frac{1}{2}, \frac{1}{2}]) \subset [-\frac{1}{2}, \frac{1}{2}]$.

$$-\frac{1}{2} < -0.012931... = g(-\frac{1}{2}) \le g(x) \le g(\frac{1}{2}) = 0.017931... < \frac{1}{2} \text{ (3 pts)}$$

Second check $|g'(x)| \le k$ for some $k \in (0,1)$, $\forall\, x \in (-\frac{1}{2}, \frac{1}{2})$.

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

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

Answer:

[Cubic]

One solution is given by (there may be others)

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

Check that $g'(p) = g''(p) = 0$ and $g^{(3)}(p) \ne 0$, and then compute

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

3

[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 $g''(p) \neq 0$, and then compute

$$\lim_{n \to \infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2!}. \quad \textbf{(5 pts)}$$

7. (10 pts) Derive Aitken's Δ^2 acceleration method.

   Hint: the starting point is to assume p_n converges to p linearly.

   Ans:

   Start from the assumption

   $$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}.$$

   See derivation in the textbook. Partial credits for partial results.

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

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

   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 orders.

   Ans:

   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

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

   where α is a 2×2 matrix and I is the identity matrix. One could choose

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

   (up to here = 5 pts)

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

   $$\mathbf{x}_* \approx (-4.882452175e - 03, \ 7.404885005e - 03)^t. \ \textbf{(10 pts)}$$